C.1

A STATISTICAL CLASSIFICATION OF BREAST CANCER PATIENTS BY DEGREE OF NODAL METASTASES

by

SANDRA LEE WILSON
B.S., Stanford University, 1966

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in the Department
of
Mathematics

We accept this thesis as conforming to the
required standard

THE UNIVERSITY OF BRITISH COLUMBIA

May 1977

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 representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Mathematics

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date June 14, 1977

ABSTRACT

Recently the traditional primary method of treatment for breast carcinoma — the Halsted radical mastectomy — has been challenged. It is felt by some people that other methods may be more appropriate for certain women. Quality of life and the patient's preferences are being considered in addition to the strictly medical aspects of the problem. One procedure that attempts to increase the quality of life for certain women is the selective biopsy. Women who are proven to have lymph node metastases at the biopsy are spared a mastectomy and treated by radiation since surgery cannot remove all of the cancer.

A study was undertaken at the British Columbia Cancer Institute of selective biopsy patients diagnosed between 1955 and 1963 in order to assess the procedure in British Columbia. After studying survival for selective biopsy patients and others, it was concluded that the procedure should continue to be recommended. Since only 14% of the patients now referred to BCCI have had a selective biopsy, I decided to try to find a statistical method for assessing the probability of nodal metastases. The problem is one of statistical classification. The literature on the theory of several statistical models was reviewed. Two models were chosen for the problem: linear discriminant analysis and logistic regression. The classification procedure most often used is discriminant analysis. However, the linear discriminant model assumes a normal distribution and

and common covariance matrix for the vector of observations. Medical data is often non-normal and even discrete. The logistic probability model works well with such data. Both models were then used to study the selective biopsy problem.

The patients of the BCCI study were used as a training set to estimate the parameters of the discriminant function and the logistic probability function. Then each estimated function was used to classify the patients as a measure of the goodness of fit of the models. The logistic regression correctly classified slightly more of the patients than the discriminant analysis did. Because of the iterative nature of the logistic regression, the execution time for the logistic regression was longer than for discriminant analysis, but not beyond practical limits.

The variables that were significant in the statistical analyses could be used to help the physician make a clinical assessment of the lymph nodes of a woman with breast carcinoma. The variables indicate areas where further research would be useful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

The data for this work was provided by the British Columbia Cancer Institute. Particular thanks go to Dr. Glen Crawford of BCCI for arranging for me to work with the data and the Statistics Department of BCCI for help with the medical files.

Thanks also go to my committee members for their comments and help in the preparation of this work. Those committee members are

Dr. Stanley J. Nash
Department of Mathematics

Dr. Brenda J. Morrison
Department of Health Care and Epidemiology

Dr. S. James Press
Faculty of Commerce

Finally, thanks must go to my husband for his encouragement and support throughout my quest for a degree.

Chapter 1

INTRODUCTION

Cancer is one of the most universally feared dieases known to man. Not only does it too often kill, it kills slowly and usually with pain and suffering. The treatments for this dread disease sometimes seem worse than the disease itself: amputations of parts of the body, radiation to kill cells (both cancerous and normal), and chemicals that poison cells. For a woman, breast cancer usually holds the greatest fear because, in addition to the physical damage done by the disease and treatments, there is often great emotional damage. The North American and European cultures have put such emphasis upon a woman's breasts in defining her worth as a woman that deformity or loss of a breast is an emotional blow that can cripple a woman. In addition, breast cancer "is the single largest cause of death from cancer among women in the United States and Canada" [34, p. 334].

In the treatment of cancer there are presently three types of treatment.[1] The oldest and most often used as an initial therapy is surgery. If the cancer is completely removed, then the disease is no

---

[1] A new form of treatment called immunotherapy is being tried experimentally but is not generally available and so is not discussed here.

longer a problem. However, cancer does not confine itself to neat easily-excised tumors. Single cells that break off from the main mass can travel via the blood and lymphatic systems to all parts of the body and establish new colonies of cancer cells called metastases. To remove as many as possible of these cells that have broken away, cancer surgery removes wide areas of presumably normal tissue in addition to the tumor itself. This can cause major physical deformities. Even such extensive surgery is often not enough to stop all the cancer cells.

Because some cancers are inoperable (not amenable to surgery because of the size or location of the tumor), other methods of treatment are necessary. Radiation is known to kill cells, normal and abnormal alike. Radiation can reach places that surgery cannot and does not cause as much deformity. However, it, like surgery, cannot kill all the stray cells.

A systemic treatment was needed to kill the colonizing or metastatic cells. It has been found that certain drugs kill cancerous cells faster than normal cells because cancer cells have a more rapid rate of growth than normal cells. Thus, chemotherapy became another weapon in the treatment of cancer. While chemotherapy does not cause permanent physical deformities, it does cause temporary distressing side effects.

Treatment of a particular breast cancer patient can be by any one of these methods or by any combination. Too often the treatment is dictated by the physician's personal preference rather than by the circumstances of the case. Some doctors have tried methods of assessing the best treatment for the patient by taking into account the quality

of life of the patient and other possible rewards under alternative treatments. One such study was completed in the spring of 1976 at the British Columbia Cancer Institute (BCCI).

A surgical procedure called a selective biopsy was done after an initial diagnosis of breast cancer. This procedure attempted to determine whether a patient had lymph node metastases or not. Depending on the status of the lymph nodes, a course of treatment was recommended. Between the years 1955 and 1963, 557 women had a selective biopsy done and were referred to BCCI for further treatment. The medical staff at BCCI undertook a study to compare the results of different treatment methods for these patients. Some results of that study will be reported in Chapter 2 as the background for the problem to be studied here.

Definitive conclusions are not always possible from the selective biopsy because of contamination, loss of material, or incomplete dissection. Also many patients do not have the selective biopsy done. A statistical model is proposed in this paper to augment, and possibly supplant for some patients, the surgical procedure. The patients that provide the data base for this work are the same ones that were used in the study conducted at BCCI. The statistical problem of deciding whether there are nodal metastases or not is a two-group classification problem. Four models for classification of mixed (discrete and continuous) data will be discussed in Chapter 3. Two of these models — linear discrimination and logistic regression —will then be applied to the problem of classifying breast cancer patients by degree of nodal metastases.

Chapter 2

MEDICAL HISTORY OF THE PROBLEM

In 1882 Dr. William Halsted began performing the first true radical mastectomies in Baltimore, Maryland. A radical mastectomy involves the "removal of the breast, pectoral muscles, axillary lymph nodes, and associated skin and subcutaneous tissue" [17]. Surgeons quickly adopted his operation as the standard treatment for breast carcinoma. It remains the most widely used procedure today and is the standard against which other treatments are judged.

Other surgical treatments range from a lumpectomy (removal of only the tumor mass) through super-radical operations that remove even more tissues than the radical mastectomy does. These surgical procedures combined with various types of radiation and chemotherapy produce a large range of combinations of treatments. In the women to be studied here there were twelve different types of treatment combinations (see Appendix A). Only women are considered in this study because of the different factors that are thought to affect the disease in men and women.[1]

The variations in treatment reflect the preferences of the physician treating the woman in addition to the variations in the disease

---

[1] Breast carcinoma in men manifests itself in the same areas — lymph nodes, muscles, and breast tissues —however, the hormonal influences are thought to be quite different.

process. The treatment a woman gets depends more on the doctor she consults than on the state of her disease. To try to eliminate these differences caused by doctor variability, attempts have been made to set up standard treatment protocols. During the time of the study (1955-1963) radical mastectomy with post-operative radiation therapy was the treatment of choice for operable cases of breast carcinoma seen at the British Columbia Cancer Institute (BCCI). Radiation alone was considered the best treatment for inoperable cases.

The next step was to decide which cases were operable. This has been where most of the differences of opinion occurred. It is generally agreed that growth of the disease beyond the breast makes a case inoperable. Any type of metastases constitutes such growth. All researchers agree that the prognosis is poor, no matter what the treatment, if, as Haagensen says, "metastases had reached these lymph nodes at the periphery of the regional lymph node filter at the apex of the axilla and in the internal mammary chain" [22, p. 691]. Thus, it seems reasonable to consider patients with nodal metastases as having growth beyond the breast and thus being inoperable. The method of assessment of these apical and internal mammary nodes was the next problem.

Several clinical[2] assessment systems have been devised in order to try to predict the pathological findings. The system used at BCCI is the Manchester staging of breast cancer (see Appendix B). There are four clinical stages that are supposed to correspond to four pathological stages. The stage I's are early disease while the stage IV's represent

---

[2]Clinical findings are those obtained from physical examination of the patient without surgery. Pathological findings are those obtained from microscopic examination of surgically obtained tissue samples.

advanced disease for both clinical and pathological scales. Clinical III's and clinical IV's are generally considered inoperable "because the likelihood of cure by radical mastectomy is so poor that other methods will do as well or better" [22, p. 691]. Thus, we consider only clinical I's and II's as possible operable cases.

Unfortunately, the clinical staging systems have not done very well at predicting pathological staging. As Haagensen says, "clinical features alone upon which we relied for the selection of patients betrayed us. . ." [22, p. 691]. The results from BCCI, as presented in Table I are typical. For clinical I's 50% have negative nodes, while for clinical II's 49% had pathologically involved apical or internal mammary nodes.

In order to permit pathological review of the nodes before a radical mastectomy was carried out, a procedure called a selective or triple biopsy was devised. Dr. C.D. Haagensen developed and used this procedure between 1951 and 1966. His results are the only published findings of large groups using selective biopsy and surgery. With a combination of clinical staging and selective biopsy, it was hoped that a better assessment of the state of the disease could be made before any treatment, including mastectomy, was begun.

The selective biopsy is recommended for Clinical I's with inner half lesions, central lesions, or outer half masses with tumors larger than 3 cms and for all Clinical II's [11]. It begins with the original biopsy of the tumor mass in the breast. When the rush report[3] is positive

---

[3]A rush report is the report of the frozen section done while the patient is still in surgery. A permanent, or paraffin, section is done later because it is more accurate and shows greater detail.

Table I

Clinical Staging versus Pathological Staging

(489 cases)

Pathological	Clinical I	Clinical II	Clinical III & IV
I	50.2%	20.4%	11.1%
II	21.4%	30.6%	2.8%
III	0.8%	0.0%	8.3%
IV	27.6%	49.0%	77.8%
	100.0%	100.0%	100.0%
Number of Patients	257	196	36

for malignancy, the apical and internal mammary lymph nodes (areas III and IV in Figure 2.1) are also biopsied. In early use of the procedure, the tissues obtained in the second stage were also subject to a rush section and further surgery, if indicated, was undertaken immediately. However, results of the rush sections of the nodes were often inconclusive and the paraffin sections were necessary for accurate evaluation. Thus, the present procedure evolved in which the patient is returned to her room until the permanent sections are read. If the internal mammary and apical lymph nodes are negative, the woman is returned to the operating room for a radical mastectomy and later referred for radiation treatment to the supraclavicalur (another name for apical) and internal mammary areas. When any of the nodes are positive, the patient has no further surgery and is referred for radiotherapy to the breast and all node areas [22,11].

Haagensen stopped doing selective biopsies in 1967 because he felt that he had learned all that he could from them [22]. The staff of the BCCI did not feel that was an adequate reason for discontinuing a practice that offered such advantages and so they continue to recommend its use for Clinical I's and II's. Since most patients are referred to BCCI after initial surgery has been performed, the decision to do the biopsy usually remains with the attending physician. The procedure was most popular during the late 1950's when a high of 43.5% of the patients referred to BCCI had had the biopsy done. It has declined in popularity until the present time when about 18% of the patients referred have had the selective biopsy performed. Because of the increasing incidence of breast cancer and increasing referral to BCCI, the number of patients having the biopsy each year has increased despite the percentage decrease.

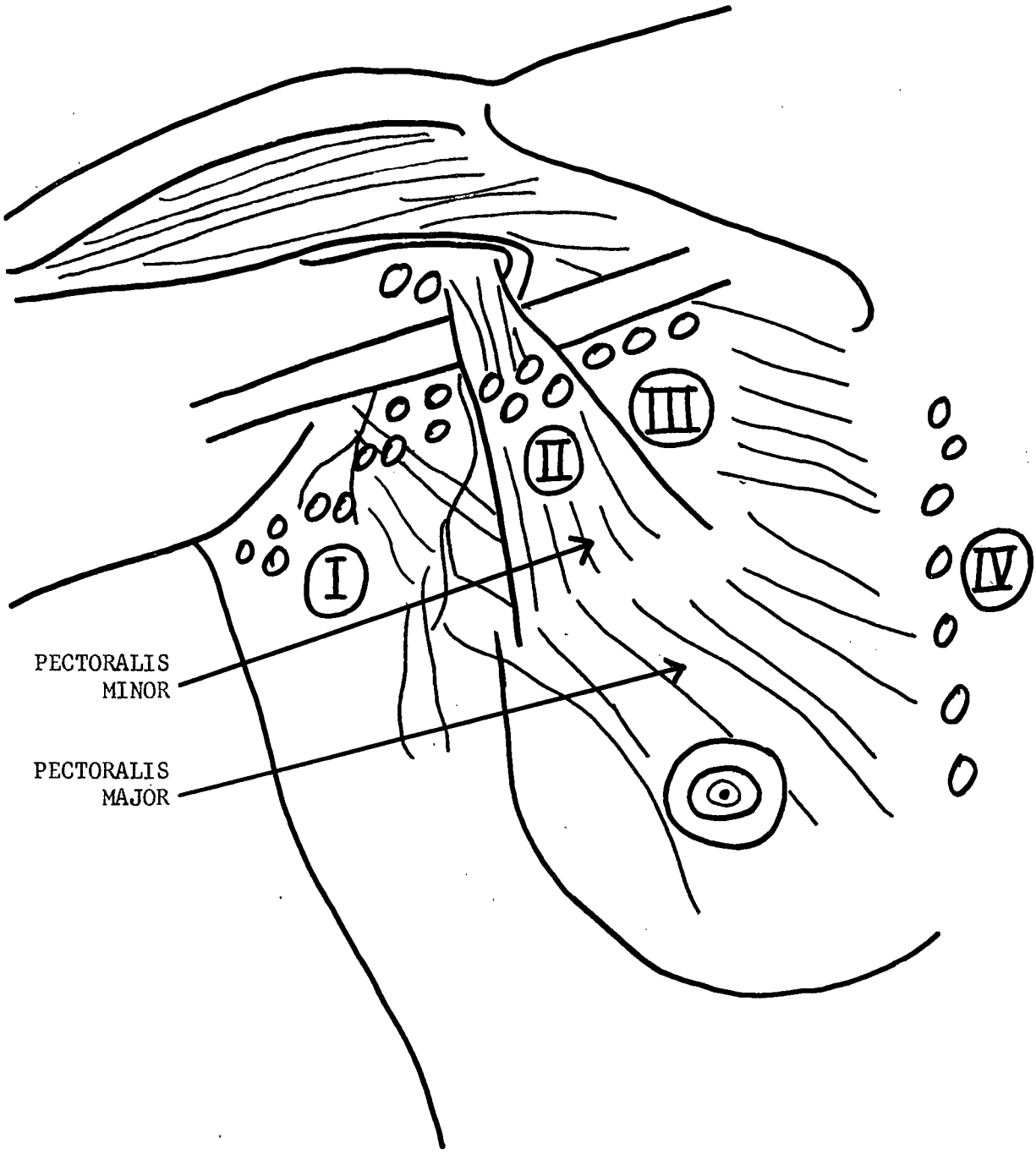


Figure 2.1. Areas of lymph node involvement in Carcinoma of the breast.

In order to assess the results of using the selective biopsy to select patients for surgery in British Columbia, a retrospective study was undertaken at BCCI of selective biopsy patients whose date of diagnosis was between 1955 and 1963. Since five and ten year survival rates are the standards for comparison in cancer therapy, the years to be studied were chosen to ensure availability of ten year survival data for all patients. A total of 557 women were referred to the BCCI after selective biopsy in the specified years. Twenty-two patients were eliminated from the study because of previous breast malignancy (14) or other systemic malignancy (8). Skin cancer and carcinoma in situ of the uterus did not constitute cause for being removed from the study. The remaining 535 women were then put into treatment groups by the method of treatment they actually received. Ideally there would have been only two groups: 1. radical mastectomy with radiation to apical and internal mammary nodes and 2. radiation to original lesion and axillary, apical, and internal mammary drainage areas. However, due to the fact that patients came from many different referring surgeons, there were twelve different treatment groups (see Appendix A for details of the groups). Only the two recommended groups (called C and A) had enough cases to give significant statistical results.

The first concern of the doctors was that the selective biopsy procedure did not harm the patient. It was known that there was little morbidity associated with the procedure. To judge whether it affected survival, all selective biopsy patients were compared to all other new breast cases for 1955 to 1963. The data are presented in Table II. To test whether the survival rates were worse for selective biopsy patients,

Table II

Survival of Selective Biopsy and All Other New Breast Cases by Clinical Stage

(1955-1963)

Selective Biopsies

Clinical Stage	Number of Cases	Alive at 5 Years	Alive at 10 Years
I	271	207	154
II	202	107	75
III	35	18	14
IV	13	2	0
Unknown	14	10	9
Total	535	344	252

Other New Cases

Clinical Stage	Number of Cases	Alive at 5 Years	Alive at 10 Years
I	529	390	304
II	266	137	88
III	133	59	34
IV	235	29	8
Unknown	34	21	14
Total	1197	636	448

a series of 2x2 contingency tables were formed for five and ten year survival. The contingency tables are presented in Table III, where the different treatments are selective biopsy or not selective biopsy.

We now wish to test whether the proportions in the two treatment groups differ significantly for each contingency table. That is, we wish to test the hypothesis that survival is independent of the treatment group. Since we must estimate the parameters, the appropriate test is a chi-square with one degree of freedom. We calculate

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \tag{2.1}$$

where f_{ij} is the (i,j)th observed cell frequency and F_{ij} is the corresponding expected cell frequency. F_{ij} is calculated as follows:

$$F_{ij} = \frac{f_{i\cdot}\, f_{\cdot j}}{f_{\cdot\cdot}} \tag{2.2}$$

(a dot indicates summation over that index)

where $f_{i\cdot}$ is the i-th row marginal total, $f_{\cdot j}$ is the j-th column marginal total, and $f_{\cdot\cdot}$ is the grand total of all cases. Only F_{11} needs to be calculated that way since all other F_{ij} are uniquely determined by F_{11} and the fixed row and column marginals. It is a property of 2x2 tables that $f_{ij} - F_{ij}$ is the same except for sign for all i,j=1,2. Thus, we get

$$\chi^2 = (f_{11} - F_{11})^2 \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{1}{F_{ij}} \tag{2.3}$$

Table III

2x2 Contingency Tables for Survival

Clinical I

	5 years			10 years		
	Alive	*Dead*		*Alive*	*Dead*	
Other	390	139	529	304	225	529
Selective	207	64	271	154	117	271
	597	203	800	458	342	800

Clinical II

	5 years			10 years		
	Alive	*Dead*		*Alive*	*Dead*	
Other	137	129	266	88	178	266
Selective	107	95	202	75	127	202
	244	224	468	163	305	468

Clinical III

	5 years			10 years		
	Alive	*Dead*		*Alive*	*Dead*	
Other	59	74	133	34	99	133
Selective	18	17	35	14	21	35
	77	91	168	48	120	168

Clinical IV

	5 years			10 years		
	Alive	*Dead*		*Alive*	*Dead*	
Other	29	206	235	8	227	235
Selective	2	11	13	0	13	13
	31	217	248	8	240	248

Clinical Unknown

	5 years			10 years		
	Alive	*Dead*		*Alive*	*Dead*	
Other	21	13	34	14	20	34
Selective	10	4	14	9	5	14
	31	17	48	23	25	48

Total

	5 years			10 years		
	Alive	*Dead*		*Alive*	*Dead*	
Other	636	561	1197	448	749	119
Selective	344	191	535	252	283	53
	980	752	1732	700	1032	172

and χ^2 is asymptotically distributed as a chi-square with one degree of freedom [44]. A correction for continuity should be added giving the final result

$$\chi_c^2 = (|f_{11} - F_{11}| - 0.5) \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{1}{F_{ij}} \quad . \tag{2.4}$$

The above approximate procedure can be used when the numbers in the tables are large (all expected frequencies are greater than five). However, when the total number of cases is less than 20 or the smallest expected frequency is less than five, it should not be used [44].

In 1935 Fisher showed that an exact test of significance based on the hypergeometric probability distribution could be made.[4] Finney *et al.* [14] calculated these probabilities and published tables of the results for up to 40 total cases. When the contingency tables in Table III involved small numbers, these exact probabilities were used for testing for significant differences.

All 2x2 tables for individual clincial stages at five and ten years showed differences that were not significant at the .05 level. Thus, we cannot reject the hypothesis that survival was the same. However, some tables had a great disparity between the numbers of patients in the two treatment groups (for example, Clinical IV Other 235, Selective 13). For that reason it was decided to test five and ten year survival for all clinical stages combined. When the chi-square test was used, it showed a significant different ($P < .01$). Thus, we would reject the hypothesis

---

[4]The exact conditional probability is

$$\frac{f_{1\cdot}!\, f_{2\cdot}!\, f_{\cdot 1}!\, f_{\cdot 2}!}{f_{\cdot\cdot}!\, f_{11}!\, f_{12}!\, f_{21}!\, f_{22}!} \, .$$

that total survival was the same. Since total survival was better for the selective biopsy group, we can conclude that selective biopsy at least did not decrease patient survival. Thus, we conclude that the selective biopsy did not harm the patients in this study.

The fact that selective biopsy patients demonstrated longer survival may be a result of being included in an "experimental" group. They may have been followed more closely and thus had metastases (local and distant) treated earlier. It is also possible that a larger proportion of these patients received radiotherapy as part of their treatment and consequently recurrences were delayed.

The next step in assessing the selective biopsy as used in British Columbia was to compare survival rates to published results and to compare survival rates between groups. It was concluded that the local survival rates were not significantly different ($P > .05$) from the published results of Haagensen [22] (Table IV). Five and ten year survivals were then compared for treatment groups A (no mastectomy, standard radiation) and C (radical mastectomy, standard radiation). The results are shown in Table V. The differences were significant at the .05 level for both time periods. This was expected since the A cases were known to be advanced cases while the C cases were known to be early cases. The final comparison was for five and ten year survivals for patients known to have nodal metastases and treated by mastectomy - group C - versus those known to have nodal metastases and treated by irradiation - group A - (Table VI). The number of patients with nodal disease and treated by mastectomy was small (18), but not too small to attempt to draw some conclusions. It was found that at five years there was no significant difference at the .05

Table IV

Contingency Table for 10 Year Survival for BCCI Patients and Haagensen's Patients

	Alive	*Dead*		*Survival*
Haagensen	550	526	1076	51.1%
BCCI	254	279	533	47.7%
	804	805	1609	

Table V

Contingency Tables for 5 and 10 Year Survival for Treatment Groups A and C

5 years

	Alive	*Dead*		*Survival*
A	61	85	146	41.8%
C	241	75	316	76.3%
	302	160	462	

10 years

	Alive	*Dead*		*Survival*
A	26	120	146	17.8%
	198	118	316	62.7%
	224	238	462	

Table VI

Contingency Tables for 5 and 10 Year Survival for Patients with Nodal Metastases and in Treatment Group A versus Treatment Group C

5 years

	Alive	*Dead*	
A	57	84	141
C	10	8	18
	67	92	159

10 years

	Alive	*Dead*	
A	21	120	141
C	8	10	18
	29	130	159

level of significance. However, at ten years we could reject the hypothesis of no significant differences ($P < .05$). It was not known what selection factors were used to select the cases with positive nodes for surgery (a typical problem with any retrospective study). The recommendation was made by the BCCI medical staff to continue use of the selective biopsy.

After the data were collected for the analysis that comprises the main body of this work, more information was available on survival. For some patients a 22 year history of survival was available and thirteen year survival information was available for all patients. I decided to do an actuarial survival study to see what had happened in the years between five and ten and in the years after ten. The Biomed program BMD11S-Life Tables and Survival Rate was used to produce actuarial survival rates. The results are shown in Figure 2.2 through 2.5. Figure 2.2 shows the overall survival for selective biopsy patients. The results at five and ten years compare favorably with standard survival results [7]:

Standard - 5 years	60%
Selective biopsy - 5 years	64.6%
Standard - 10 years	20%
Selective biopsy - 10 years	47.7%

There is also a 28% survival for 22 years which is encouraging.

A comparison of Figures 2.3 and 2.4 shows the curves to be almost identical — the survival for group A and the survival for positive nodes are nearly the same and the survival for group C and the survival for negative nodes are the same. Again that is the expected result since A cases were supposed to be chosen because of the presence of positive nodes while C cases are supposed to have negative nodes. Figure 2.5 shows the

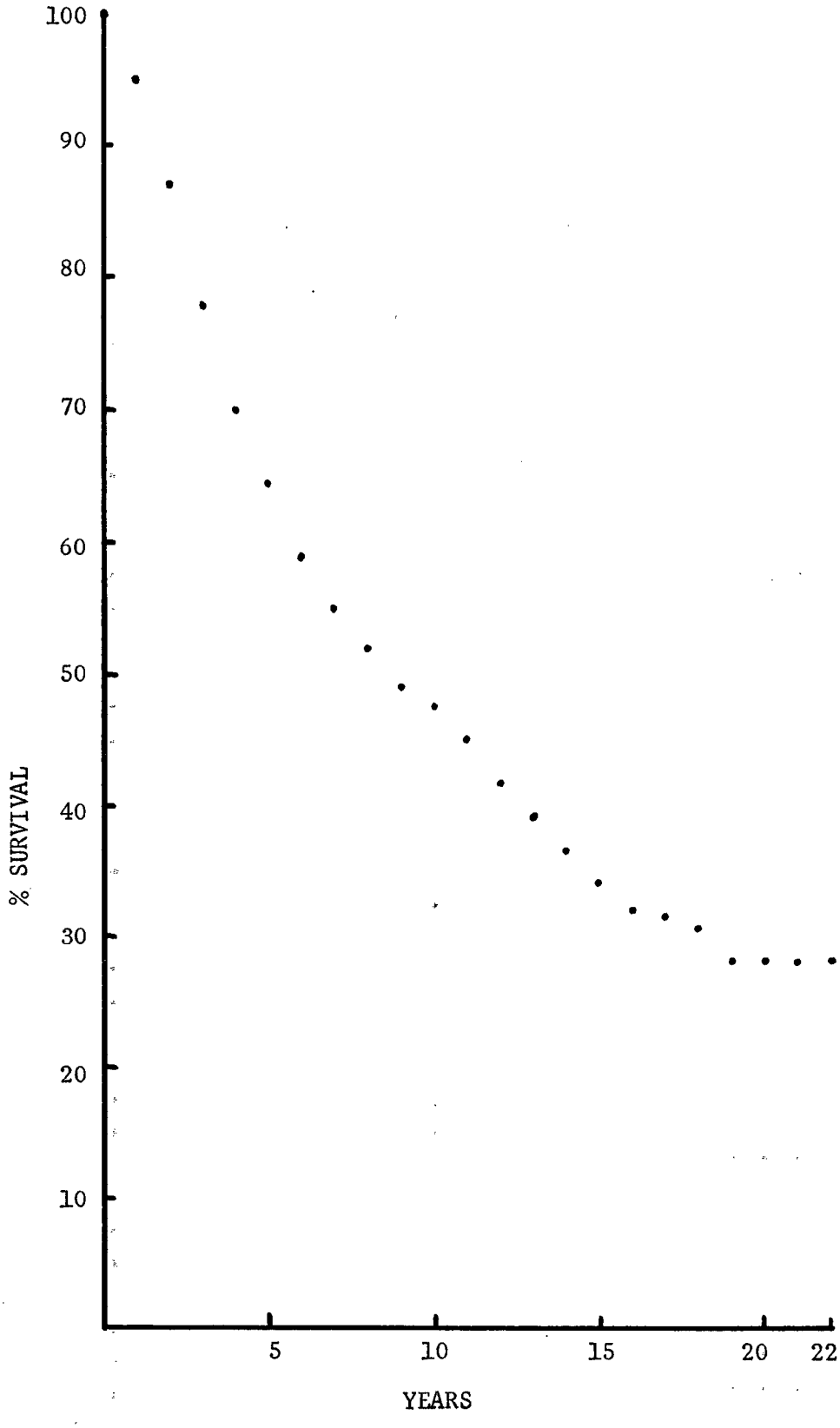


Figure 2.2. Actuarial survival of all 535 cases.

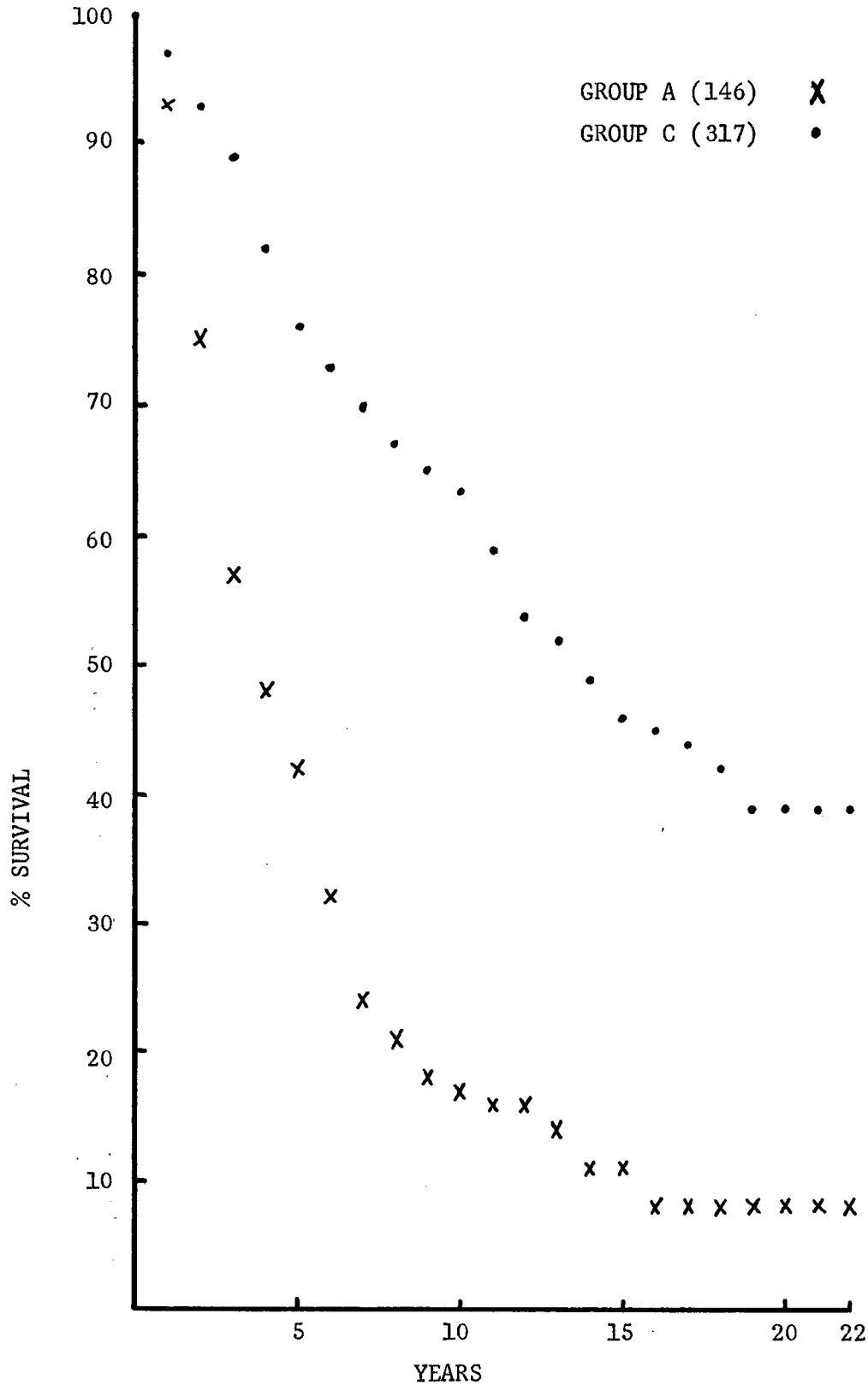


Figure 2.3. Actuarial survival of patients in treatment group A versus patients in treatment group C.

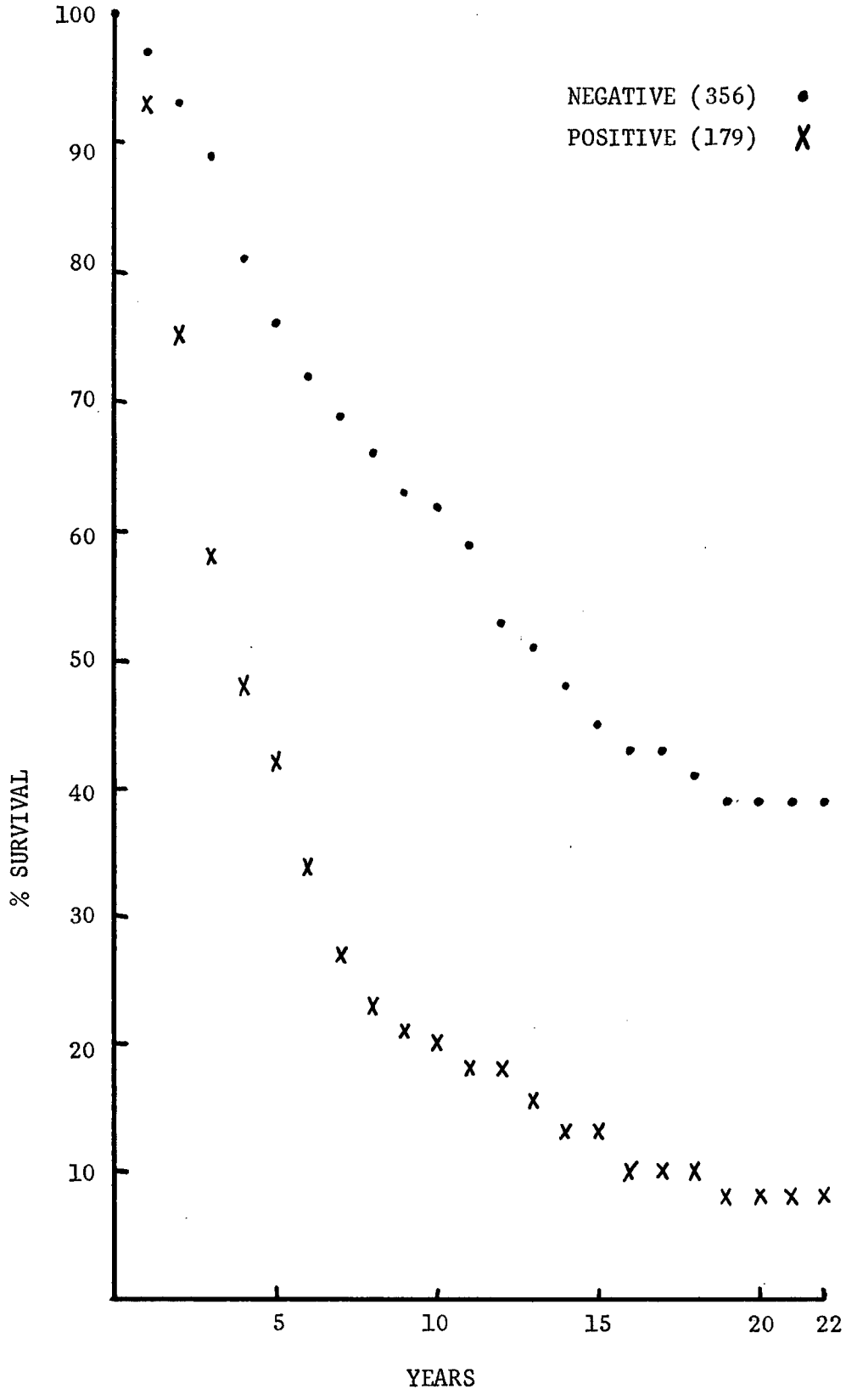


Figure 2.4. Actuarial survival of patients with positive nodes versus patients with negative nodes.

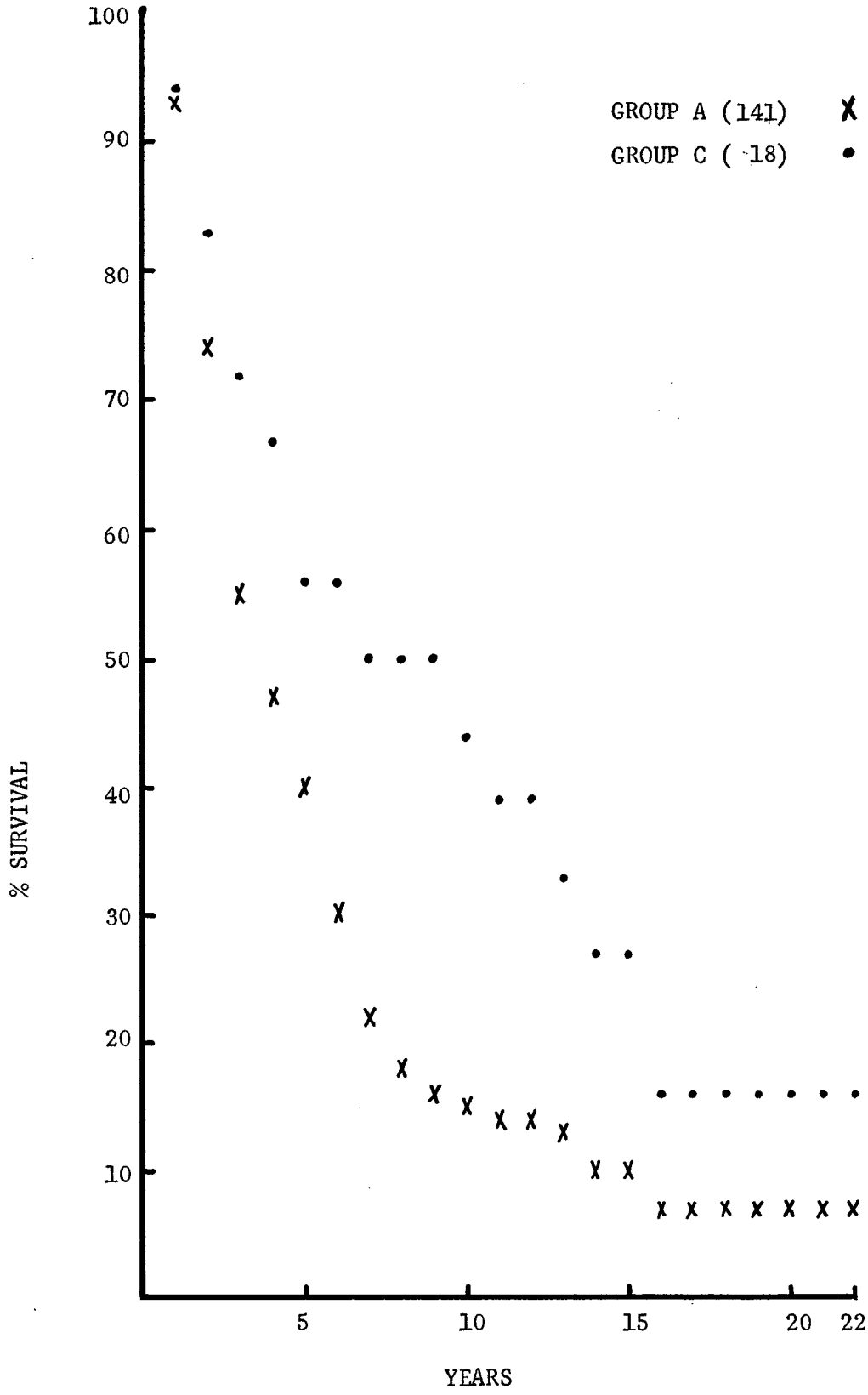


Figure 2.5. Actuarial survival of patients with positive nodes and in treatment group C versus patients with positive nodes and in treatment group A.

survival curves by treatment group for patients with positive nodes. The curves are close together through five years. Then between six and twelve years they are divergent. After twelve years they approach each other again. The Biomed survival program also calculates a t-test for the differences between groups each year. The results of these t-tests appear in Table VII. These results confirm that there is a significant difference ($P < .05$) in the years six through twelve only. It is not clear just what significance this has for the treatment decision problem. More cases with positive nodes and treatment by mastectomy need to be studied with that question in mind.

Another question of concern about the selective biopsy was the local recurrence rate — that is, recurrence of the disease in the breast and associated lymph drainage areas. Not all the data were available on recurrence for the BCCI study. It was known that local recurrences were more of a problem in study group A. However, most local recurrences were adequately controlled. More work remains to be done on the question of local recurrence.

To complete the assessment of the use of selective biopsy, the medical staff have been asking questions about the quality of life for the patients. They feel that sparing a woman with advanced breast carcinoma the mutilation of having a breast removed is giving her a better quality of life. However, the quality of life also has a time component to it. Mueller and Jeffries studied the questions of rate of dying and causes of death in breast carcinoma and concluded

Table VII

T-test of Survival Differences for Positive Nodes in Group A and Group C

Year	T-statistic	Degrees of Freedom	Tabled $t_{.975}$ (df)
1	-0.26	157	1.9763
2	-1.00	157	1.9763
3	-1.49	157	1.9763
4	-1.67	157	1.9763
5	-1.22	157	1.9763
6	-2.03*	157	1.9763
7	-2.28*	157	1.9763
8	-2.58*	157	1.9763
9	-2.83*	157	1.9763
10	-2.44*	157	1.9763
11	-2.08*	157	1.9763
12	-2.08*	157	1.9763
13	-1.79	157	1.9763
14	-1.60	155	1.9765
15	-1.60	155	1.9765
16	-0.86	118	1.980
17	-0.86	118	1.980
18	-0.86	118	1.980
19	-0.86	118	1.980
20	-0.86	118	1.980
21	-0.86	118	1.980
22	-0.86	118	1.980

*significant

> Breast cancer treatment should: a) Treat the cancer only when and where it is known to exist; b) Not be proposed as a means of influencing either time of death or cause of death.
>
> Measurements of quality of life should be established and should constitute the only realistic objective of treatment [34, p. 339].

Thus, the conclusion of the BCCI study was that the selective biopsy should be recommended for patients with Clinical stage I or Clinical stage II disease [11].

Since BCCI is a referral agency for all of British Columbia and the Yukon, most of the range of stages were well represented in the study. The group of patients was deficient in very early cases of Clinical stage I which had received surgery and then were not referred for further treatment. Presumably these would all have had negative nodes since evidence of any nodal metastases in the surgical specimen would be cause for referral. The study could also be deficient in very advanced stages where the patient would be beyond any treatment. Since that stage of disease would never be recommended for selective biopsy, we need not worry about lack of such cases.

After completing the statistical analysis of the above study for BCCI, I became interested in trying to find a statistical model that could classify patients into positive or negative nodes when surgical results were not available. Since 82% of the patients now being referred to BCCI have not had a selective biopsy, it could be useful for helping to decide on the best treatment for these patients. It could also be used with those patients for whom the selective biopsy was inconclusive.

Chapter 3

REVIEW OF STATISTICAL MODELS

The medical diagnosis problem presented in the previous chapter can be considered as a statistical classification or prediction problem. Given a vector of observable variables for a patient, we wish to predict which group that patient belongs to (positive or negative nodes). Several different models have been suggested for this problem. Four of the models will be discussed here: Fisher's linear discriminant function, multiway contingency tables, Krzanowski's location model, and the logistic probability model. Discussion will include the assumptions of the models, parametric estimation methods, problems in using the models, and availability of computer routines.

3.1 Fisher's Linear Discriminant

In 1936, R.A. Fisher proposed a linear discriminant function to classify a p dimensional vector $\underline{X}$ into one of two known multivariate normal populations, given that the observation was from one of the two and that they had the same covariance matrix. We assume that

$$\underline{X} \sim N_p(\underline{\mu}_1, \Sigma) \text{ with probability } q_1$$

and

$$\underline{X} \sim N_p(\underline{\mu}_0, \Sigma) \text{ with probability } q_0$$

where $q_1 + q_0 = 1$ and Σ is the common covariance matrix. The linear discriminant function is

$$U(\underline{X}) = \beta_0 + \underline{\beta}' \underline{X}, \tag{3.1}$$

where

$$\beta_0 = \ln \frac{q_1}{q_0} - \tfrac{1}{2} \sum_{i=1}^{p} \beta_i(\mu_{i1} + \mu_{i0}) \tag{3.2}$$

and

$$\underline{\beta}' = (\underline{\mu}_1 - \underline{\mu}_0)' \Sigma^{-1} \tag{3.3}$$

so that

$$\beta_i = \sum_{j=1}^{p} (\mu_{i1} - \mu_{i0})\sigma^{ij} \text{ for } i=1,\cdots,p \tag{3.4}$$

with $\Sigma^{-1} = (\sigma^{ij})$.

If $U(\underline{X}) \geq 0$, $\underline{X}$ is assigned to population 1, otherwise, $\underline{X}$ is assigned to population 0.[1]

The goal of the parameter estimation procedure is to minimize expected total misclassification cost. Often the costs for misclassification are quite different for the two populations (death versus further testing in a medical diagnosis problem). One can include these costs in the model and then minimize the expected total cost of misclassification.

[1] In actual medical practice, those individuals for whom $U(\underline{X})$ is zero or near zero would not be classified without further investigation. A two-stage procedure which allows further observation of borderline cases is discussed later in this chapter in the section on variable reduction.

$C(h|k)$ is defined as the cost of misclassifying an individual to group h when that individual is a member of group k. The expected misclassification cost for group k is $q_k C(h|k)$ and the total expected misclassification cost is

$$\sum_k q_k C(h|k). \tag{3.5}$$

Replacing q_k by $q_k C(h|k)$, the linear discriminant model becomes

$$U(\underline{X}) = \beta_0 + \underline{\beta}' \cdot X \tag{3.6}$$

with

$$\beta_0 = \ln r - \tfrac{1}{2} \sum_{i=1}^{p} \beta_i(\mu_{i1} + \mu_{i0}), \tag{3.7}$$

$$r = \frac{q_1\ C(0|1)}{q_0\ C(1|0)}, \tag{3.8}$$

and

$$\underline{\beta}' = (\underline{\mu}_1 - \underline{\mu}_0)'\ \Sigma^{-1} \tag{3.9}$$

so that

$$\beta_i = \sum_{j=1}^{p} (\mu_{i1} - \mu_{i0})\sigma^{ij}\ . \tag{3.10}$$

Again $\underline{X}$ is classified into group 1 when $U(\underline{X}) \geq 0$, and into group 0 otherwise.

When the parameters of the populations are unknown, they must be estimated. We shall assume that the sampling is random from the mixture of populations so that the sampling mixture approximates the population mixture. When there is a low incidence of one population, a two-sample procedure (separate samples for the different groups) may be more

appropriate [1]. However, the parameter estimation would be different for that case. Since the patients in the selective biopsy study were sampled from the mixture of populations, we will not consider the separate sample situation.

Let n_h = number of observations from group h, h=0,1, and x_{iht} = the i-th characteristic of the t-th individual in the h-th group. Then

$$\bar{x}_{ih} = n_h^{-1} \sum_{t=1}^{n_h} x_{iht}, \qquad h=0,1, \tag{3.11}$$

$$s_{ij,h} = (n_h - 1)^{-1} \sum_{t=1}^{n_h} (X_{iht} - \overline{X}_{ih})(X_{jht} - (\overline{X}_{jh}), \qquad h=0,1, \tag{3.12}$$

and

$$\hat{\sigma}_{ij} = \frac{(n_1 - 1)S_{ij,1} + (n_0 - 1)S_{ij,0}}{n_1 + n_0 - 2} \tag{3.13}$$

To estimate the population proportions we use the sample proportions. Thus,

$$\hat{q}_h = \frac{n_h}{n_1 + n_0}, \quad h=0,1,$$

so that

$$\frac{\hat{q}_1\, C(0|1)}{\hat{q}_0\, C(1|0)} = \frac{n_1\, C(0|1)}{n_0\, C(1|0)}. \tag{3.14}$$

The population means are estimated by the sample means: $\hat{\mu}_h = \underline{\overline{X}}_h$. Let $\hat{\sigma}_{ij}$ be the (i,j)th element of $\hat{\Sigma}$ and $\hat{\sigma}^{ij}$ be the (i,j)th element of $\hat{\Sigma}^{-1}$. Thus, the estimated function is

$$\hat{U}(\underline{X}) = \hat{\beta}_0 + \underline{\hat{\beta}}' \cdot X \tag{3.15}$$

where

$$\hat{\beta}_0 = \ln \frac{n_1 C(0|1)}{n_0 C(1|0)} - \tfrac{1}{2} \sum_{i=1}^{p} \hat{\beta}_i (\hat{X}_{i1} + \hat{X}_{i0}) \tag{3.16}$$

and

$$\hat{\beta}_i = \sum_{j=1}^{p} (\overline{X}_{i1} - \overline{X}_{i0}) \hat{\sigma}^{ij}. \tag{3.17}$$

If $\hat{U}(\underline{X}) \geq 0$, $\underline{X}$ is assigned to population 1, otherwise, $\underline{X}$ is assigned to population 0. Rewriting $\hat{U}(\underline{X})$ to clarify the estimation problems we get

$$\hat{U}(\underline{X}) = \ln \frac{n_1 C(0|1)}{n_0 C(1|0)} + (\underline{\overline{X}}_1 - \underline{\overline{X}}_0)' \hat{\Sigma}^{-1} [\underline{X} - \tfrac{1}{2}(\underline{\overline{X}}_1 + \underline{\overline{X}}_2)]. \tag{3.18}$$

We see from (3.18) that in order to estimate the linear discriminant function we must estimate $\underline{\mu}_1$, $\underline{\mu}_0$, and Σ. Unless some simplifying assumptions are made about Σ (for example, independence of variates), the estimation problem can become quite substantial.

Departures of the data from normality are a cause for concern with this model. Although little has been done to show robustness of the linear discriminant functions, many practical applications proceed with linear discrimination after stating that the data are non-normal or even discrete [see for example, 47]. This problem will be of great concern here because it is known that the medical data are non-normal and often not even continuous.

An attractive feature of the linear discriminant model for applications is the widespread availability of computer routines for estimating the function and classifying observations. The discriminant analysis is based on a linear regression and so is easily accessible. The availability of the computer program encourages the user to ignore the departures from the model assumptions for ease of computation.

3.2 Contingency Tables

Each individual has a set of attributes describing him. When all the data are discrete, all individuals with the same set of attributes are counted and that count is put into the appropriate cell of a contingency table. The structure of the table for two variables is a rectangular array with columns corresponding to levels of one variable and rows corresponding to levels of the other variable.

The simplest contingency table is a 2 x 2 table. There are two levels of attribute A and two levels of attribute B. The model would appear as below:[2]

		B: 1	B: 2	
A	1	p_{11}	p_{12}	$p_{1\cdot}$
	2	p_{21}	p_{22}	$p_{2\cdot}$
		$p_{\cdot 1}$	$p_{\cdot 2}$	1

where p_{ij} is the (i,j)th cell probability,

$$p_{i\cdot} = \sum_{k=1}^{2} p_{ik}, \qquad i=i,2, \tag{3.19}$$

$$p_{\cdot j} = \sum_{k=1}^{2} p_{kj}, \qquad j=1,2, \tag{3.20}$$

and

$$\sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1. \tag{3.21}$$

The $p_{i\cdot}$ and $p_{\cdot j}$ are the marginal probabilities. The model assumptions are that all categories or contingencies are included (the probabilities

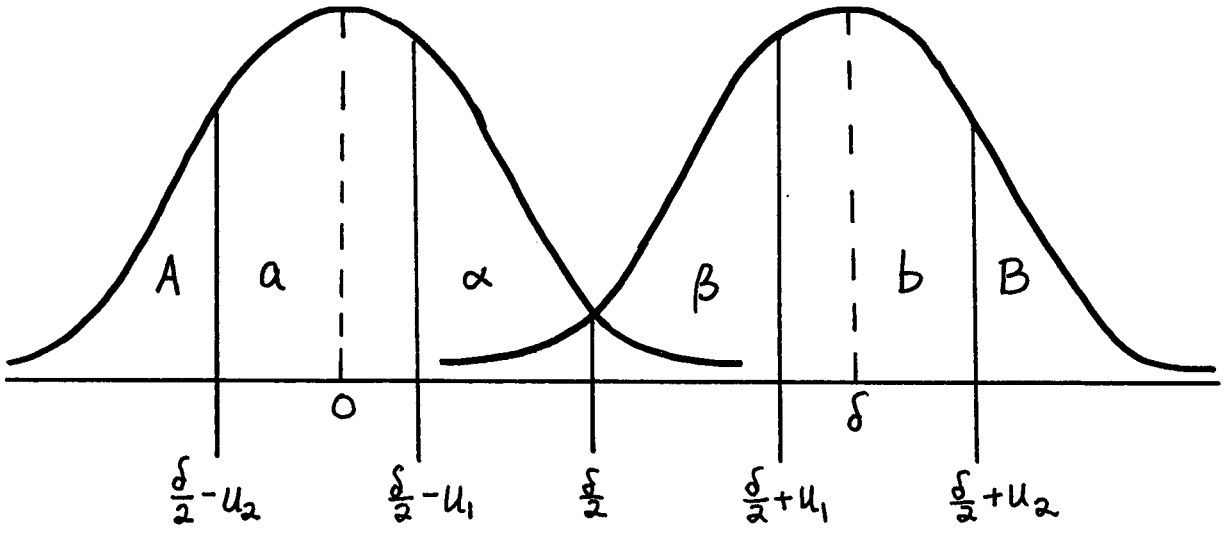
[2]The model could also be written in terms of the expected frequencies $m_{ij} = Np_{ij}$.

sum to one) and that all variables are discrete (or have been made discrete). This model and equations (3.19), (3.20), and (3.21) easily generalize to higher dimensions.

The general model assumptions remain the same: all variables are discrete and the probabilities in all tables sum to one. In higher way contingency tables, however, one usually makes some simplifying assumptions about interaction terms to make the problem more manageable. One common simplification is to assume that bivariate interactions are allowed, but higher order interactions are not.

Cochran and Hopkins [8] suggested that when there is a mixture of continuous and qualitative varibles, all the continuous variables should be made qualitative. They concluded that the optimal partition would be into six states as shown in Figure 3.1. When the variables are all qualitative, the problem has been reduced to analysing a p-way contingency table. The question with this approach is how much information is lost. Cochran and Hopkins felt the loss of information was not significant, however, many others have found the loss unacceptable and sought ways of utilizing the full information.

Estimation of the cell frequencies (or cell probabilities) in a contingency table is simple for small tables and large data sets, but can become quite complicated, if not impossible, for larger tables and moderate data sets. The number of individuals observed for each cell is enumerated and that count is the observed frequency or estimated frequency. The problems arise when there are many cells to be estimated and not very many data points. A small 3 x 5 x 7 table having fixed marginals has 48 parameters to be estimated. Another problem is empty cells. The frequency



U_1 and U_2 are calculated from the data.

Figure 3.1. Cochran and Hopkins categories for partitioning continuous variables.

method can only estimate an empty cell as zero, while it is quite likely that a different sample would show the cell to be non-empty.

It has been suggested by many authors [17, 24, and 35 for example] that log-linear models are appropriate for analysing contingency tables. Log-linear models fit the contingency table model assumptions, while solving the estimation problems discussed above. In many cases, empty cells can be estimated as non-zero with log-linear models. Also with a few simplifying assumptions (high order interactions are zero for example), there are many fewer parameters to estimate so that for a given data set size the estimated parameters of the log-linear model will be based on more observations per parameter. A more complete discussion of log-linear models will be deferred to Section 3.4.

Another problem with high dimension contingency table analysis has been the general unavailability of computer routines to do the analysis. Recently UCLA's Biomed package has included a program for analysis of a multiway table. Greater availability of this program will encourage more use of contingency table analysis with high dimension problems. The availability of computer routines will not alleviate the problems of large numbers of parameters to be estimated and loss of information with partitioned continuous variables.

3.3 Krzanowski Location Model

In order to use all the information available when there is a mixture of continuous and discrete data W.J. Krzanowski proposed a likelihood ratio classification rule based on the location model [27]. In the

location model $\underline{X} = \begin{pmatrix} \underline{Y} \\ \underline{Z} \end{pmatrix}$ where $\underline{Y}$: qxl is the vector of continuous variables and $\underline{Z}$: pxl is the vector of binary variables.[3] Thus, each distinct pattern of $\underline{Z}$ defines a multinomial cell with $\underline{Z}$ being in cell

$$m = 1 + \sum_{j=1}^{p} z_j \cdot 2^{(j-1)} .$$

It is assumed that $\underline{Y} \sim N_p(\underline{\mu}_i^{(m)}, \Sigma)$ in cell m, where Σ is the common covariance matrix. It is also assumed that the probability of an observation in cell m is p_m. The optimal allocation rule then becomes: allocate to group 1 if

$$U(\underline{Y}) = \left[\underline{\mu}_1^{(m)} - \underline{\mu}_0^{(m)}\right]' \Sigma^{-1} \left[\underline{y} - \tfrac{1}{2}(\underline{\mu}_1^{(m)} + \underline{\mu}_0^{(m)})\right] + \ln \frac{p_{0m}}{p_{1m}} \qquad (3.22)$$

is ≥ 0 and to group 0 otherwise.

> The optimum rule derived from the location model thus leads effectively to a different linear discriminant for each of the multinomial cells, with cutoff points determined in each case by the discrete component of the model [27, p. 783].

Thus, this is a model that acknowledges the different types of variables, but it is unduly complicated in the number of functions produced. The location model seems to be of theoretical interest but of little practical use at this time.

Krzanowski does, however, suggest that if the data are not to be treated by his method but rather by fitting to a model containing only

---

[3] A discrete variable with n levels can be transformed into n-1 binary indicator variables, so we consider only binary variables here.

one type of variable (either continuous or discrete, but not both), then it is better to consider them all as continuous rather than to partition the continuous variables. Thus, for a mixture of continuous and discrete variables, he preferred Fisher's linear discriminant to p-way contingency table analysis.

3.4 Logistic Regression

A simpler model for continuous and discrete variables is the logistic probability model. It allows both continuous and discrete variables without loss of information and the estimators have several desirable properties.

Let $P(\underline{X}_i)$ be the posterior probability that an individual with explanatory variable values $\underline{X}_i' = (X_{i1},\cdots,X_{im})$ has the disease (belongs to group 1). Then

$$P(\underline{X}_i) = \left(1 + \exp(-\beta_0 - \underline{\beta}' \underline{X}_i)\right)^{-1}, \tag{3.23}$$

(a prime on a vector indicates the transpose of the vector) where β_0 and $\underline{\beta}$ are the logistic coefficients. The expression in (3.23) is the multivariate logistic function. In the medical context, (3.23) is a good formulation because "in the light of present medical knowlege a reasonable assumption is that P follows a symmetric sigmoid curve" [49, p. 168]. Cox [in 10] showed that the logistic function is appropriate for several different types of distributions of the explanatory variables: multivariate normal, binary, and mixed. In general, (3.23) holds for any

variables whose distributions are in the exponential family; that is, those with density functions of the form

$$f(\underline{X}) = g(\underline{\theta})\, h(\underline{X}) \exp\{T(\underline{\theta})\, \underline{X}\}. \tag{3.24}$$

Assuming that the distribution of $\underline{X}$ is described by (3.24) is such a mild restriction that we can ignore it for all practical purposes. A multivariate generalization of (3.23) can be made quite easily for problems with more than two groups. The generalization would allow division into 2^k classes where P is a vector with k components.

The logistic probability model has several desirable properties in addition to its general applicability for different distributions. Cox [10] showed that (3.23) has associated with it the simple sufficient statistics

$$\underline{t} = \sum_{i=1}^{n} \underline{X}_i y_i \tag{3.25}$$

where the y_i are indicator variables corresponding to group membership (0-1). The maximum likelihood estimators are functions of the sufficient statistics. Thus, from the Rao-Blackwell Theroem, we expect to get smaller mean squared error using the maximum likelihood estimators than using estimators that are not functions of sufficient statistics. In addition, the logistic model has asymptotically unbiased estimators associated with it [25].

Several different procedures have been suggested to estimate the parameters of the logistic function. Unfortunately, all the procedures are, by necessity, iterative. Walker and Duncan [49] derive the normal

equations through a least squares procedure with estimated weights.

We let

$$P_i = P(\underline{X}_i) = \left[1 + \exp(-\underline{\beta}' \underline{X}_i)\right]^{-1} \tag{3.26}$$

be the probability of the i-th individual of the sample having the disease.

Therefore,

$$y_i = P_i + \varepsilon_i = f(\underline{\beta}, \underline{X}_i) + \varepsilon_i. \tag{3.27}$$

Thus, the n x (m+1) matrix of independent variables for the sample is

$$X = \begin{bmatrix} X_{10} & X_{11} & \cdots & X_{1m} \\ X_{20} & X_{21} & \cdots & X_{2m} \\ \cdots & & & \\ X_{n0} & X_{n1} & \cdots & X_{nm} \end{bmatrix}$$

and $\underline{y}' = (y_1, y_2, \cdots, y_n)$. By application of weighted iterative non-linear least-squares procedures to (3.27) with a diagonal weight matrix W (the inverse of the covariance matrix of the vector $\underline{\varepsilon}$), they conclude that the normal equations are

$$X' W^{-1} X \underline{\beta} = X' W^{-1} \underline{y}. \tag{3.28}$$

Thus, they conclude that the estimators are

$$\hat{\underline{\beta}} = (X' W^{-1} X)^{-1} X' W^{-1} \underline{y} , \tag{3.29}$$

which we shall see later are the same as the maximum likelihood estimators. Walker and Duncan suggest an iterative solution of (3.29) by the Newton-Raphson method with initial estimates obtained by fitting a linear discriminant function.

Kalman [49] proposed another recursive estimation procedure that claimed more rapid convergence and because of the rapid convergence, the need for good initial estimates is relaxed. The procedure updates the estimates with the addition of each new individual. The estimate of $\underline{\beta}$ based on the first k individuals is $\hat{\underline{\beta}}_k$. Then

$$V_k = \text{Var}(\hat{\underline{\beta}}_k) = (X_k' W_k^{-1} X_k)^{-1} \tag{3.30}$$

where X_k is the matrix of k individuals' observations and W_k is the covariance matrix for k individuals. Let $\underline{X}_{k+1}$ be the vector of observations for the (k+1)st individual,

$$w_{k+1} = \frac{1}{\hat{P}_{k+1} \hat{Q}_{k+1}} , \tag{3.31}$$

and

$$\hat{P}_{k+1} = \hat{P}_{k+1|k} = \left(1 + \exp(-\hat{\underline{\beta}}_k' \underline{X}_{k+1})\right)^{-1} = 1-\hat{Q}_{k+1} . \tag{3.32}$$

Therefore,

$$V_{k+1} = V_k - V_k \underline{X}_{k+1} c_{k+1} \underline{X}_{k+1}' V_k \tag{3.33}$$

and

$$c_{k+1} = (w_{k+1} + \underline{X}_{k+1}' V_k \underline{X}_{k+1})^{-1} . \tag{3.34}$$

Finally, the recursive formula for the estimator of $\underline{\beta}$ is

$$\hat{\underline{\beta}}_{k+1} = \hat{\underline{\beta}}_k + V_k \underline{X}_{k+1} c_{k+1} w_{k+1}(y_{k+1} - \hat{P}_{k+1}) \tag{3.35}$$

where y_{k+1} takes on the values 1 and 0 as the (k+1)st individual does or does not have the disease.

The problem of intial estimates is quite simple now. Let V_0 and $\underline{b}_0$ be any prior estimates of the variance and $\underline{\beta}$. The estimates V_k and $\hat{\underline{\beta}}_k$ are found using V_0, $\underline{b}_0$, and the first k data items. Then V_0 and $\underline{b}_0$ are eliminated from the formula by the following:

$$V_k^* = \left(V_k^{-1} - V_0^{-1}\right)^{-1} , \tag{3.36}$$

and

$$\hat{\underline{\beta}}_k^* = V_k\left(V_k^{-1} \hat{\underline{\beta}}_k - V_0^{-1} \underline{b}_0\right) . \tag{3.37}$$

The remaining m-k items, V_k^*, and $\hat{\underline{\beta}}_k^*$ are then used in the recursive process to get the final estimate of $\underline{\beta}$.

A third method is maximum likelihood estimation. The maximum likelihood equations are fairly simple to derive for the logistic model. Let P_s be the posterior probability of disease from equation (3.23) for

the s-th individual. Also let $y_s = 0$ if the s-th individual does not have disease and $y_s = 1$ if that individual has disease. Then the likelihood is

$$L = \prod_{s=1}^{n} P_s^{y_s} \left(1 - p_s\right)^{1-y_s} \tag{3.38}$$

and the natural logarithm of the likelihood is

$$\ln L = \sum_{s=1}^{n} y_s \ln P_s + \sum_{s=1}^{n} (1 - y_s) \ln(1 - P_s). \tag{3.39}$$

Taking partial derivatives of (3.39), we get the maximum likelihood equations[4]

$$\frac{\partial \ln L}{\partial \beta_0} = \sum_{s=1}^{n} y_s - \sum_{s=1}^{n} P_s = 0 \tag{3.40}$$

and

$$\frac{\partial \ln L}{\partial \beta_i} = \sum_{s=1}^{n} y_s X_{is} - \sum_{s=1}^{n} X_{is} P_s = 0 \qquad i=1,\cdots,p. \tag{3.41}$$

Equation (3.40) assures us that the expected number of cases will be equal to the observed number of cases. That is another desirable property of the maximum likelihood estimates.

---

[4] To take the partial derivatives we use the facts that $\frac{\partial \ln P_s}{\partial \beta_0} = 1-P_s$, $\frac{\partial \ln(1-P_s)}{\partial \beta_0} = -P_s$, $\frac{\partial \ln P_s}{\partial \beta_i} = X_{is}(1-P_s)$ and $\frac{\partial \ln(1-P_s)}{\partial \beta_i} = -X_{is} P_s$.

The maximum likelihood equations are most often fitted by the Newton-Raphson algorithm. It is an iterative gradient algorithm. If p_k is an estimate of P, $k \geq 0$, and f is the logistic function to be estimated, then the new estimate p_{k+1} is

$$p_{k+1} = p_k - (d^2 f)^{-1} (df). \tag{3.42}$$

With this formulation there can be a problem of divergence when the initial estimates are not close to the true values. Thus, Haberman [in 24] added a factor $\alpha(k)$ to equation (3.42) to prevent divergence in such cases. If reasonable care is taken in choosing the starting values, divergence is not a large problem in most applications so we will not complicate (3.42) with the α term.

Several different types of starting estimators have been suggested in the literature. Linear discriminant function estimators have often been used as starting values. Other possible initial estimators are conditional estimators and reverse Taylor series approximations [35]. Conditional estimators are obtained by maximizing the conditional likelihood (conditional on the explanatory variables). Reverse Taylor series approximations arise from the logistic function, equation (3.23). Expanding about $\underline{X} = \overline{\underline{X}}$ in a Taylor series, one gets

$$P(\underline{X}) = \left\{ \frac{1}{1 + \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}})} - \frac{\underline{\beta}'\overline{\underline{X}} \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}})}{[1 + \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}})]^2} \right\}$$

$$+ \left\{ \frac{\underline{\beta}' \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}}}{[1 + \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}}]^2} \right\} \underline{X} + R(\underline{X}), \tag{3.43}$$

where $R(\underline{X})$ denotes a remainder containing terms of the order, $0[\underline{X} - \overline{\underline{X}})'(\underline{X} - \overline{\underline{X}}]$. Neglecting the remainder, we can interpret this as the linear function $A + \underline{B}'\underline{X}$ where

$$A = \frac{1}{1 + \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}})} - \underline{B}'\overline{\underline{X}} \tag{3.44}$$

and

$$\underline{B} = \frac{\underline{\beta} \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}})}{[1 + \exp(-\beta_0 - \underline{\beta}'\overline{\underline{X}})]^2} . \tag{3.45}$$

Solving one gets

$$\hat{\underline{\beta}} = \frac{\underline{B}}{(A + \underline{B}'\overline{\underline{X}})\ (1 - A - \underline{B}'\overline{\underline{X}})} \tag{3.46}$$

and

$$\hat{\beta}_0 = -\hat{\underline{\beta}}'\overline{\underline{X}} - \ln\left(\frac{1}{A + \underline{B}'\overline{\underline{X}}} - 1\right) \tag{3.47}$$

as the reverse Taylor series approximations.

Computer routines to find the maximum likelihood estimators for the logistic model or logistic regression are not so readily available as those for linear discrimination. However, they are becoming more accessible. An example of one such program is listed in the work by Nerlove and Press [35]. Like most logistic regression programs, it uses the Newton-Raphson algorithm to find the MLE's. A disadvantage of this

routine is the necessity for an additional user-written program to calculate the probabilities and classifications.

Because the logistic formulation provides a probability of being in group 1 rather than just a classification, one can tell which individuals are quite likely to be correctly classified (probability near 0 or 1) and which individuals are near the boundary (probability near 0.5) and thus are likely to be incorrect. The results could easily be used to form three groups: 1) those in group 0, 2) those in group 1, and 3) those in the middle region for whom more investigation should be carried out before classification. This is a particularly desirable characteristic for a medical diagnosis problem. Some tests are expensive, while others are inexpensive. If a patient can be classified (diagnosed) on the basis of inexpensive tests only, it is desirable for the patient and medical staff. However, if the first tests are inconclusive, the more costly tests are available to help resolve the question. Thus, we can think of the logistic regression as giving us a two-stage procedure.

3.5 Comparison of Linear Discrimination and Logistic Regression

For theoretical and practical considerations outlined above, I chose to concentrate on only two of the classification models: linear discrimination and logistic regression. Linear discrimination was chosen because of its widespread use in published studies despite violations of the model assumptions and because of easily accessible computer routines. Logistic regression by use of the Newton-Raphson algorithm was selected because of the good fit of the medical data to the model assumptions, the desirable features of the logistic estimators, and the availability of a

computer program. In another work [41] S.J. Press and I compared logistic regression and discriminant analysis. Theoretical arguments were presented for and against the use of logistic regression as opposed to discriminant analysis for classification and regression of qualitative variables on explanatory variables. Empirical results for some non-normal classification problems were reported.

A theoretical comparison of linear discrimination and logistic regression under different conditions is the next concern of this work. When the data are multivariate normal with equal covariance matrices, the model assumptions of both models are satisfied. One would expect the linear discriminant to be better in this case because its model assumptions are satisfied, it has a closed form and it is non-iterative. Also for the normal case, both types of estimators are asymptotically unbiased.[25].

Efron [12] investigated the asymptotic efficiency for each model under the assumption of normality of the data. He calculated the asymptotic relative efficiency (ARE) of logistic regression to linear discrimination. The ARE is given by

$$\text{ARE} = \frac{1 + \Delta^2 \, q_1 \, q_0}{(2\pi)^{\frac{1}{2}}} \exp\left(\frac{-\Delta^2}{8}\right) \int_{-\infty}^{\infty} \frac{\exp(-x^2/2)\, dx}{q_1 \exp(\Delta x/2) + q_0 \exp(-\Delta x/2)} \tag{3.48}$$

where

$$\Delta = \left[(\underline{\mu}_1 - \underline{\mu}_0)' \, \Sigma^{-1} \, (\underline{\mu}_1 - \underline{\mu}_0)\right]^{\frac{1}{2}},$$

which is the square root of the Mahalanobis distance. Table VIII gives some sample results for the asymptotic relative efficiency when $q_1 = q_0 =$

Table VIII

Asymptotic Relative Efficiency of Logistic Regression to Linear Discrimination ($q_1 = q_0 = .05$)

Δ	0	.5	1	1.5	2	2.5	3	3.5	4
ARE	1.000	1.000	.995	.968	.899	.786	.641	.486	.343

0.5, the most favorable situation for logistic regression. The logistic regression is less efficient as n goes to infinity because the linear discriminant is based on the full maximum likelihood estimators for β_0 and $\underline{\beta}$ while logistic regression is a conditional maximum likelihood estimation procedure. Thus, when the data are multivariate normal, the linear discriminant approach is to be preferred on theoretical grounds.

Since "the multivariate normal assumption is unlikely to be satisfied in applications, even approximately,..." [25, p. 125], this seems the more important situation to consider. When the normality assumption is violated, logistic regression is theoretically more robust. The logistic probability model is valid for the exponential family, equation (3.24). The question of relative efficiencies in this case has not been investigated, but one would expect the logistic estimators to be more efficient. It was shown above that when (3.23) holds, there are sufficient statistics for $\underline{\beta}$. The maximum likelihood estimators are functions of the sufficient statistics, but the discriminant function estimators are not. Thus, by the Rao-Blackwell theorem, the logistic estimators have a smaller expected mean square error than the discriminant function estimators.

Under non-normal conditions, the logistic maximum likelihood estimates are consistent (asymptotically unbiased). For the discriminant

function procedure, Halperin, Blackwelder, and Verter showed that for large samples and one attribute variable

$$\hat{\beta} \to (P_1 - P_0)/(q_1P_1Q_1 + q_0P_0Q_0) \tag{3.49}$$

and

$$\hat{\beta}_0 \to \ln\left[\frac{q_1P_1 + q_0P_0}{q_1Q_1 + q_0Q_0}\right] - \frac{q_1}{2}\left[\frac{P_1}{q_1P_1 + q_0P_0} + \frac{Q_1}{q_1Q_1 + q_0Q_0}\right]$$

$$\cdot \left[\frac{P_1 - P_0}{q_1P_1Q_1 + q_0P_0Q_0}\right] \tag{3.50}$$

where $$P_1 = \left(1 + \exp(-\beta_0 - \beta)\right)^{-1} = 1 - Q_1 \tag{3.51}$$

and $$P_0 = \left(1 + \exp(-\beta_0)\right)^{-1} = 1 - Q_0. \tag{3.52}$$

When $q_1 = 0.5$ the estimates are nearly unbiased, but for other values of q_1 and q_0 the discrepancies can be quite large. Equations (3.49), (3.50), (3.51), and (3.52) can be easily generalized to more than one attribute variable [25]. The authors continued the analysis to show that

> a) β_i which are zero will tend to be estimated as zero for large samples by the method of maximum likelihood, but not necessarily by the discrimination method;
>
> b) if any β_i are non-zero they will tend to be estimated as non-zero by either method, but the discriminant function approach will give asymptotically biased estimates for those β_i and for (β_0). . . [25, p. 152].

Finally, if there are two or more β_i that are non-zero, the discriminant function estimates of the β_j that are zero will not converge to zero in

general. Consequently, one could be led to believe that certain factors are significant when in reality they are not. Empirical studies have shown that it does indeed happen in non-trivial cases. Thus, when the data follow the exponential family but are not normal with equal covariance matrices, logistic regression is preferred on theoretical grounds.

In addition to the theoretical aspects of the comparison, some practical aspects must be considered. Since logistic regression is an iterative procedure, its estimated parameters are more complicated to calculate and thus computation is more time consuming. Halperin, Blackwelder, and Verter [25] found that logistic regression took longer by a factor ranging from 1.3 to 2. For a problem with 50 individuals and 5 independent variables, I found [in 41] that logistic regression took 1.4 times longer. The time problem gets worse as the number of variables and number of observations increase. Other things being equal, doubling the number of observation doubles the time (see Table IX for some sample times). The time can increase even more quickly than indicated in the table because the time depends too on other factors such as starting values of the coefficients, divergence of the algorithm, and covariances between the various coefficients. One would conclude that execution times can be much longer for logistic regression, but not beyond acceptable limits for large-scale computer installations.

3.6 Variable Reduction

In an exploratory data analysis of retrospective medical studies, the statistician will often have a large number of variables available. In order to make the problem more manageable and make the underlying biologic process clearer, it is desirable to reduce the dimension of

Table IX

Execution Times for Logistic Regression

(using an IBM 360-65 from [35]

Independent Variables	Observations	CPU Time for Execution
1	89	6
6	225	36
7	886	60

the problem. A number of criteria have been suggested to achieve the reduction. Three types will be discussed below: 1) checking all possible subsets, 2) a stepwise procedure, and 3) a two-stage non-iterative procedure.

McCabe [29] proposes that one check all possible subsets of the variables to find the optimal subset. This procedure has the advantage of considering all possible combinations and interactions of variables. Let the within sum of cross-products matrix be denoted by $W = (w_{ij})$ where

$$w_{ij} = \sum_{k=1}^{2} \sum_{s=1}^{n_h} \left(X_{ihs} - X_{ih\cdot}\right)\left(X_{jhs} - X_{jh\cdot}\right) \tag{3.53}$$

(a subscript replaced by a dot indicates the variable was averaged over that index)

and let the total sum of cross-products matrix be $T = (t_{ij})$ where

$$t_{ij} = \sum_{h=1}^{2} \sum_{s=1}^{n_h} \left(X_{ihs} - X_{i\cdot\cdot}\right)\left(X_{jhs} - X_{j\cdot\cdot}\right). \tag{3.54}$$

One of the standard multivariate analysis of variance tests for equality of the means uses

$$U = |W| \; / \; |T| \; . \tag{3.55}$$

That is essentially the ratio of the estimated generalized variance within to the estimated generalized variance total. Clearly, $0 \leq U \leq 1$ and small values indicate good discrimination. McCabe uses U as a descriptive statistic to show discrimination potential. Given two subsets with an equal number of variables, the subset corresponding to the smaller U

is the preferred subset. The process is similar to calculating squared multiple correlation coefficients in regression.

Theoretically, the examination of all possible subsets is the best procedure, but in practice it becomes quite lengthy for even moderate problems. To compare all subsets of p variables, U must be calculated from (3.53), (3.54), and (3.55) for $2^p - 1$ possible subsets. A CDC 6500 computer required five minutes of CPU time to find the subsets for 20 variables [29] and each added variable doubles the time required. Thus, while theoretically appealing, the examination of all subsets is not practical for a problem of the size considered in this work.

A second type of variable reduction scheme is a stepwise procedure. Individual variables are considered for inclusion or exclusion iteratively until a prespecified criterion has been achieved. The procedure adds at each step the variable which reduces the residual sum of squares as much as possible. That is, it uses a stepwise linear regression and the variable added is the one which maximally reduces the remaining unexplained variance about regression. An F-ratio is calculated from the within groups and total covariance matrices. Each variable is considered for inclusion depending on its relation to the already included variables only. Thus, all interactions are not studied as they were in the U-ratio scheme. An advantage of the stepwise procedure is the relatively large number of variables it can handle in a reasonable time. Computer routines such as the Biomed Stepwise Discriminant allow 80 variables and are quite efficient. With the virtual memory of large-scale computers, the number of variables could be increased beyond 80 with ease and without unreasonable increases in time. Although theoretically less desirable than the previous

procedure, the stepwise procedure is practically appealing for large numbers of variables.

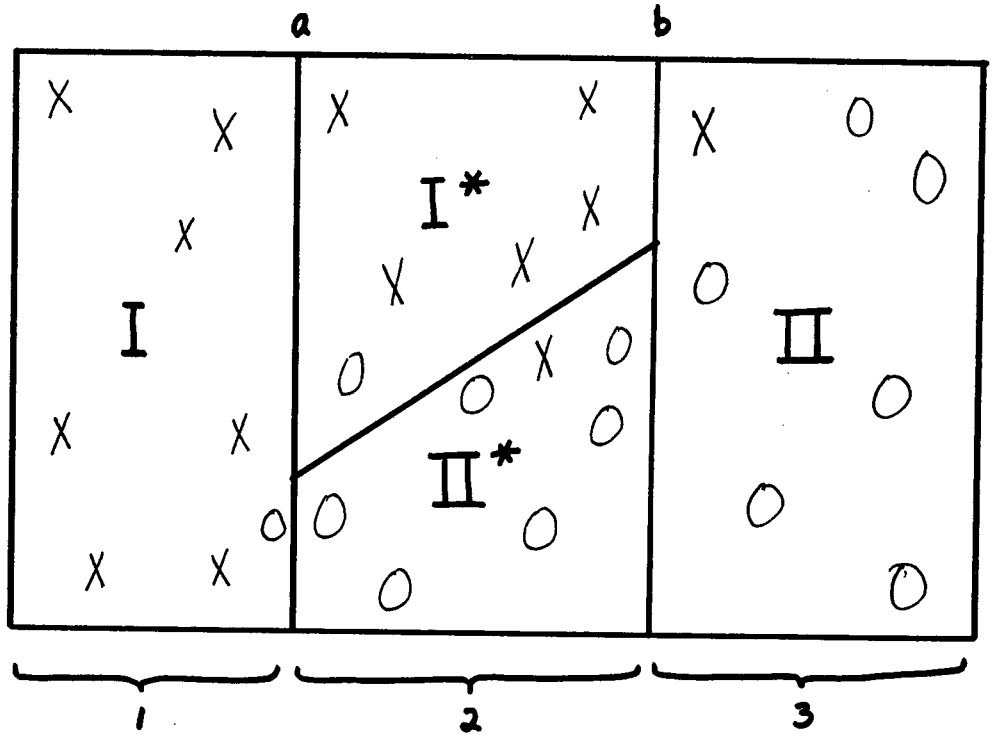
A third approach, suggested by Zielezny [51], reduces the number of variables by formalizing the medical diagnosis process. The doctor looks at an inexpensive, easily collected set of variables; if that provides a clear-cut decision, then there is no need to observe further. If it does not provide a clear-cut decision, then the doctor continues his investigation. This is especially useful when there are sets of expensive and inexpensive variables.

Let $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix}$ where $\underline{X}_j$ has k_j components and $\underline{X}_1$ corresponds to the inexpensive variables. The $\underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ are partitioned similarly. The $U(\underline{X})$ of equation (3.1) is also partitioned as

$$U(\underline{X}) = \begin{pmatrix} U_1(\underline{X}_1) \\ U_2(\underline{X}_2) \end{pmatrix} .$$

Then the two-stage rule is as follows: For a,b, $-\infty \leq a \leq b \leq \infty$, classify to group 0 if 1) $U_1(\underline{X}_1) \leq a$ or 2) $a < U_1(\underline{X}_1) \leq b$ *and* $U(\underline{X}) < 0$; otherwise, classify to group 1. The geometry of the partition is presented in Figure 3.2. That is, if $U_1(\underline{X}_1)$ is high or low an observation can be classified on the basis of $\underline{X}_1$ only. If $U_1(\underline{X}_1)$ is in the middle range, then measurements must be taken on the second subset. The parameters a and b are chosen to minimize the total cost.

Although it has not been done, it seems that this procedure could easily be generalized to p-1 stages so that variables are observed singly. One would then have a sequential decision process. Further investigation



The first stage partitions into 3 regions.

The second stage partitions the middle region.

Figure 3.2. The geometry of the two-stage procedure for two groups.

along these lines would be interesting, but will not be attempted here. However, a variation of the two-stage decision process can be constructed from the logistic regression results. More will be said about such a procedure in Chapter 5.

The three types of variable reduction reviewed in this section all appear useful based on theoretical considerations. When practical aspects are considered, the stepwise discriminant analysis is superior. It has also been shown [25] that variables shown to be significant by stepwise discrimination include the ones that are significant in logistic regression. The subset of variables may include non-significant ones also as explained above. Thus, a subset of variables chosen by stepwise discrimination is appropriate as a reduced set of variables for logistic regression.

3.7 Conclusions

We can conclude from our review of statistical models for classification that logistic regression should discriminate better than linear discrimination, for the medical problem of this work. The time should be greater for the logistic regression, but should not be a major problem. In order to reduce the dimension of the problem, stepwise discriminant analysis should be done. That is, stepwise discriminant analysis is appropriate for exploratory work and logistic regression is appropriate for final parameter estimation and classification with non-normal data.

Chapter 4

DATA COLLECTION

4.1 Selecting Variables to be Observed

The first problem in a statistical analysis is deciding what variables are to be observed. Because my prior knowledge of the problem was limited, external sources of information were consulted. The medical staff at BCCI were the first source of information about factors that influence the growth and spread of breast cancer. Their knowledge of the disease process included the general knowledge of breast cancer literature and their own personal experience treating the disease. The second source of facts about the disease process was the published medical reports on breast cancer. No previous work was available on predicting the spread of breast cancer to the regional lymph nodes. Because of the lack of knowledge about the factors influencing lymph node metastases, the variables chosen were those known to affect the risk of originally developing breast cancer and those known to affect the prognosis of the disease.

Factors known to influence the risk of breast cancer[1] can be divided into four types: 1) endocrine, 2) genetic, 3) immunologic, and

[1] The information about risk factors is generally agreed upon by medical authorities. The interested reader is referred to [9], [31], and [36] for more on the epidemiology of breast cancer.

4) other. Estrogens and ovarian activity have a great influence on breast cancer. Studies [see 36 for example] have shown that early menarche and late menopause (implying a longer than average period of estrogen production) increase the risk of breast cancer. Pregnancy and lactation are known to alter the hormonal balance of a woman's body, but studies have provided conflicting results about their influence on breast cancer risk. Early first parity[2] decreases the risk. Also the more children, the more the risk decreases. However, this may be accounted for by the fact that women with a large number of children would have begun having them at an early age. Finally, recent studies have produced conflicting results about the relationship between oral estrogen (birth control pill) use and breast cancer incidence. Thus, it was decided to observe all available information relating to the patients' ovarian activity, menstrual history, pregnancy and lactation history, and estrogen therapy.

Other areas of concern in endocrine function are the adrenal and thyroid glands. Adrenal dysfunction is known to contribute to carcinogenesis. One type of treatment for breast cancer metastases is an adrenalectomy. It is also known that there is a high incidence of thyroid disease in women with carcinoma of the breast. After thyroidectomy there is a low incidence of breast cancer. Thus, endocrine dysfunction seems to have an important influence on breast cancer. It was decided to observe all variables associated with endocrine imbalance, including especially thyroid and adrenal dysfunction.

---

[2]Parity is the condition of a woman with respect to having produced viable children regardless of whether the child was alive at birth.

The second type of factors includes genetic or hereditary factors. It is well-known that there is a familial predisposition for breast cancer. The incidence is consistently higher for relatives of breast cancer patients and it is inherited through either parental line. Complete information on relatives with breast cancer was an important area to observe. In addition to familial differences in risk, epidemiological studies [31] have shown racial differences. The lowest risk is among Oriental women and the highest risk is among Caucasian women. The influence of genetic factors is, however, complicated by environmental factors. Oriental women who move to North America from Asia increase their risk until several generations after immigration, their risk is close to Caucasian women's risk. On the other hand, non-white women generally have histologically more malignant tumors. Hence there is a greater likelihood of positive axillary nodes. Thus, it was decided to observe all genetic and environmental factors that were available.

The third set of factors involves the body's immune system. Although this may be the most important set of factors influencing the spread of the disease, it is the most difficult to observe. There is no good measure of the host-tumor relationship. It is known that various immune deficient states are associated with a very much greater risk of breast cancer. Multicentricity and bilaterality of the disease are clear indications of the inadequacy of the body's defenses. Seasonal variations in incidence have been noted (May to September is a period of lower incidence) which correspond to observed seasonal variations in response to other diseases. Perhaps the closest one can come to measuring the immune status of the body is the lymphocytic count. Since lymphocytes

are the body's main defense mechanism against disease, their activity indicates the level of resistance to disease processes. It is interesting to note that either a high or nil lymphocytic reaction is better than a moderate reaction. A high lymphocyte count means that the body has a good defense, while no lymphocytic reaction means that the response has not yet been triggered. A low to moderate level of lymphocytes indicates reaction that has been triggered but is inadequate to the task. It was concluded that lymphocytic reaction and season were to be observed. Also a history of diseases indicating immune deficiency, such as herpes zoster, was to be observed.

Other factors that influence risk of breast carcinoma are quite varied. Trauma to the breast is present in about 11 percent of breast cancer patients and an unknown percentage of normal women. Socio-economic level seems to play some part, although there is evidence that this is related to diet and lifestyle. Stomach cancer incidence is correlated with breast cancer incidence. Since stomach cancer is influenced by diet, doctors have suggested a link between diet and cancer of the breast. On the other hand, lower socioeconomic levels do not have as good access to medical care and so delay seeking treatment which allows the cancer to infiltrate the body more. Other systemic illnesses and previous benign breast disease increase the risk of breast cancer, too. As with many other diseases, the psychological state of the patient can be crucial to treatment response and survival. Thus, it was decided that all available information on the above factors was to be collected.

A final area that seemed promising was the pathological diagnosis. Breast cancer is more than one disease. Many different histological

types are combined in the label of breast carcinoma. Different histological types were known to have varying incidences among different groups of women. Table X shows the results of one study [31]. Consequently, all the pathological findings available in the medical charts were to be included as variables.

In addition to the factors that influence risk of developing breast cancer, the clinical state of the disease at diagnosis is an important indicator of possible lymph node metastases. Different states have different prognoses. A short survival associated with a particular symptom indicates a more advanced stage of disease or more virulent disease. Both of these conditions have a greater chance of nodal metastases. Size of the primary tumor can be indicative of disease state. A large mass would be seen when there is a long delay between onset of symptoms and diagnosis. The longer the tumor has been growing, the better established it is in the body. Thus, metastases are more likely. Large tumor masses also are seen in fast-growing cancers. A carcinoma that grows quickly has a good chance of having spread elsewhere. Skin involvement and clincial involvement of the axilla also indicate spread of disease and increased likelihood of lymph node metastases. Consequently, all variables pertaining to the state of the disease were to be observed.

After the review of the literature on risk factors and prognostic factors, a sample of 50 charts of patients who had not had a selective biopsy but who were initially treated at BCCI during the years 1955 to 1963 were reviewed. The charts were read to see how much of the information outlined above was recorded in patient charts and in what format the information appeared. A few additional variables that had no obvious

Table X

Comparison of Factors by Histologic Type

Histologic Type	Infiltrating duct	Infiltrating lobular	Medullary	Colloid	Comedo Carcinomas	Papillary
% of total	78.1%	8.7%	4.3%	2.6%	4.6%	1.2%
Average Age (years)	50.7	53.8	49.0	49.7	48.6	51.9
Location	all	standard	More in upper half of breast	standard	Mostly subareolar	More in lower half of breast
Nodal involvement	60%	60%	44%	32%	32%	17%

connection to the lymph node metastases problem but were available on most charts were included in the list of variables to be observed.

4.2 Data Collection

From the list of factors mentioned in the previous section, a data collection form and coding instructions were devised. The data collection form and accompanying instructions are in Appendix C. The format for data collection was suggested by previous studies undertaken at BCCI. Over a period of three months forms were completed by me for the 557 selective biopsy patients diagnosed between 1955 and 1963 and treated at BCCI. If some information was unavailable for a patient, a missing-data code was used.

After the data collection was completed, the twenty-two patients with previous breast or systemic malignancy were eliminated since the disease process could be obscured by the other primary tumors. Of the remaining 535 patients, three more were dropped from the study because of very poor information. The physicians recording the history reported that two of the women were bad historians or suspect reporters. The third woman removed from the study spoke only German and the interpreter was a young boy unfamiliar with medical terminology and biological processes. An additional three cases were discarded because expert pathological review was unable to provide a disease status for the lymph nodes. Thus, the final sample size was 529 patients.

4.3 Variables Selected for Analysis

After the sample had been pared to its final size, the variables to be included in the analysis had to be selected. Since the exploratory analysis was to be carried out by stepwise discriminant analysis, no more than 80 variables (the program limit) could be included. The posterior knowledge after data collection became the prior knowledge used for variable selection.

Three types of variables had been observed: 1) continuous, 2) binary, and 3) categorical. The binary and continuous variables did not have to be transformed prior to analysis. Some categorical variables were changed after data collection because certain categories had too few observations. The variables were collapsed by combining categories until the remaining categories were large enough for analysis. An example was the categorical variable race. All categories other than Caucasian were very small. Categories were combined to form the binary variable for Caucasian and non-Caucasian. If a categorical variable was an ordered categorical variable, it could be used without transformation. Unordered categorical variables had to be transformed into binary indicator variables. Each binary variable corresponds to one of the first k-1 categories. Thus, an individual in category i, $1 \leq i \leq k-1$, would have zeroes for each of the indicator variables except variable i which would be one. If an individual was in category k, the k-1 variables would all be zero. Only k-1 variables are used since k variables would be linearly dependent.

After the above transformations were made, there were 112 variables. To accommodate that number of variables, the discriminant analysis could have been used twice and the results combined. However, some variables

were not included in the analysis for various reasons. After the survival analysis of Chapter 2 was completed, the survival data were eliminated. Some variables had zero variance and so provided no information. Those were discarded. Several variables that had seemed promising before work began had so few respondents that they also were discarded. Through these means, enough variables were dropped to leave 80 variables for analysis. A list of the variables observed, but not used in the analysis appears in Appendix D.

The 80 variables chosen for analysis will be defined here. In the list that follows, a name is assigned to each variable and a definition of the variable appears.

General information:

ID-BCCI number-a six digit number for patient identification (not used in analysis, but used for case identification).

DIAMON-Month of initial diagnosis or treatment (also called the anniversary).

AGE-Age (at last birthday) at diagnosis-in years.

SOCECN-Socioeconomic level-1 if high socioeconomic level; 0 otherwise.

RACE-Racial origin-1 if Caucasian; 0 otherwise.

MARRY-Patient married-1 if patient is married at time of diagnosis; 0 otherwise.

BROKEN-Broken marriage-1 if patient had been married previously but the marriage has been dissolved by death, separation or legal action; 0 if never married or married now.

Family history:

BRMOM-Breast Cancer in patient's mother-1 if mother had developed breast cancer; 0 otherwise.

BRDAUG-Breast cancer in patient's daughter-1 if daughter had developed breast cancer; 0 otherwise.

BRSIS-Breast cancer in patient's sister-1 if sister had developed breast cancer; 0 otherwise.

BROTH-Breast cancer in other female relatives-1 if any other female relative had developed breast cancer; 0 otherwise.

CANCER-Blood relatives with cancer-number of relatives who have had cancer.

Patient's Personal History:

SMOKE-Cigarette smoker-1 if patient has a history of cigarette smoking; 0 otherwise.

Patient's Reproductive History:

REG-Regular menstrual periods-1 if patient had regular, uncomplicated menstrual periods; 0 otherwise.

DYSMEN-Dysmenorrhoea-1 if patient experienced dysmenorrhoea; 0 otherwise.

HORMON-Hormone therapy-1 if patient had a history of hormone therapy; 0 otherwise.

OTHDRG-Major drug therapy other than hormones-1 if there is a history of other drug therapy; 0 otherwise.

STATUS-Menopausal status-1 if patient is premenopausal or within 5 years of the menopause at the time of diagnosis; 0 if patient is postmenopausal.

AGEMEN-Age at menopause-age in years at menopause for postmenopausal patients; 88 for premenopausal patients.

AGEFRS-Age at first birth-patient's age in years at termination of first full-term pregnancy for para women; 0 for nullipara women.

BIRTHS-Parity-number of pregnancies carried to full term.

MISCAR-Miscarriages-number of pregnancies that did not carry to full term.

NUMNUR-Number nursed-number of lactation periods.

MONNUR-Number of months nursed-length in months of the longest lactation period.

BRFED-Patient breastfed-1 if patient was breastfed; 0 otherwise.

Patient's Illnesses:

DIABET-Diabetes-1 if patient had had diabetes; 0 otherwise.

HEART-Heart disease-1 if patient had a history of heart disease; 0 otherwise.

HYPER-Hypertension-1 if patient had a history of hypertension; 0 otherwise.

KIDNEY-Kidney disease-1 if patient had a history of kidney disease; 0 otherwise.

TB-Tuberculosis-1 if patient had a history of tuberculosis; 0 otherwise.

ANEMIA-Anemia-1 if patient had a history of anemia; 0 otherwise.

PNEUM-Pneumonia-1 if patient had a history of pneumonia or other serious lung diseases, excluding tuberculosis; 0 otherwise.

ALLERG-Allergy-1 if patient had a history of allergies; 0 otherwise.

THYROD-Thyroid disease-1 if patient had a history of thyroid dysfunction; 0 otherwise.

FIBROI-Uterine fibroids-1 if patient had a history of uterine fibroids; 0 otherwise.

DISOTH-Other disease-1 if patient had a history of other major diseases not listed above; 0 otherwise.

Patient's Surgical History (Before Present Illness)

OOPHOR-Oophorectomy-1 if patient had an oophorectomy; 0 otherwise.

HYSTER-Hysterectomy-1 if patient had a hysterectomy; 0 otherwise.

PELVIC-Other pelvic surgery-1 if patient had pelvic surgery other than those specifically listed above, especially involving the reproductive system; 0 otherwise.

GALLB-Cholecystectomy-1 if patient had gall bladder surgery; 0 otherwise.

THYSUR-Thyroidectomy-1 if thyroid had been removed; 0 otherwise.

ADRENL-Adrenalectomy-1 if the adrenal gland had been removed; 0 otherwise.

OTHSUR-Other surgery-1 if patient had a history of other surgery; 0 otherwise.

Benign Breast Ailments:

MAZODY-Mazodynia-1 if patient had mazodynia; 0 otherwise.

MASTIT-Mastitis-1 if patient had had benign breast disease during lactation; 0 otherwise.

BENIGN-Benign breast disease other than during lactation-1 if patient had a history of benign breast disease other than during lactation; 0 otherwise.

History of the Present Illness:

DURTON-Duration of symptoms-duration of symptom in months from onset of first symptom to date of diagnosis.

SYMPT1-Thickening or lump-1 if first symptom was a thickening or lump in the breast; 0 otherwise.

SYMPT2-Pain-1 if first symptom was pain in the breast or axilla; 0 otherwise.

SYMPT3-Nipple changes-1 if the first symptom was change in the contour of the nipple or discharge from the nipple; 0 otherwise.

SIZE-Size of the tumor-clinical size of the original tumor mass in cm.

LOC1-Lower inner quadrant-1 if tumor was located in the lower inner quadrant of the breast; 0 otherwise.

LOC2-Lower outer quadrant-1 if tumor was located in the lower outer quadrant of the breast; 0 otherwise.

LOC3-Upper inner quadrant-1 if tumor was located in the upper inner quadrant of the breast; 0 otherwise.

LOC4-Upper outer quadrant-1 if tumor was located in the upper outer quadrant of the breast; 0 otherwise.

LOC5-Lymph node tumor-1 if tumor was located in the axilla; 0 otherwise.

NODEPL-Palpable nodes-1 if the lymph nodes were palable in the axilla on the same side as the tumor; 0 otherwise.

SKIN-Skin involvement-1 if there was skin involvement at the site of the primary tumor; 0 otherwise.

BREAST-Breast involved-1 if the right breast was the primary site; 0 if the left breast was the primary site.

TRAUMA-Trauma to the breast-1 if the patient had a history of trauma to the involved breast; 0 otherwise.

Patient's Present Condition:

BODYSZ-Overall body size-3 ordered categories for body size.

BRSIZE-Breast size/shape-4 ordered categories for breast size.

CONDTN-Patient's general physical condition-1 if patient was in good physical condition at the time of diagnosis; 0 otherwise.

OTHILL-Other illnesses present-number of other illnesses present at the time of diagnosis.

EMOTON-Emotional problems-1 if the patient had emotional problems (other than any related to the cancer) at the time of diagnosis or shortly before; 0 otherwise.

LYMPH-Lymphocytes-percentage of lymphocytes in the blood at diagnosis.

Pathology:

NODES-Grouping variable, lymph node status-1 if apical or internal mammary lymph nodes were positive at diagnosis; 0 if lymph nodes were negative.

PATH1-Paget's disease-1 if histology was Paget's disease of the nipple; 0 otherwise.

PATH2-Noninfiltrating papillary carcinoma-1 if histology was non-infiltrating papillary carcinoma; 0 otherwise.

PATH3-Infiltrating papillary carcinoma-1 if histology was infiltrating papillary carcinoma; 0 otherwise.

PATH4-Infiltrating duct carcinoma-1 if histology was infiltrating duct carcinoma (scirrhous, adenocarcinoma); 0 otherwise.

PATH5-Colloid carcinoma-1 if histology was colloid carcinoma; 0 otherwise.

PATH6-Medullary carcinoma-1 if histology was medullary carcinoma; 0 otherwise.

PATH7 - In situ lobular carcinoma-1 if histology was in situ lobular carcinoma; 0 otherwise.

PATH8-Infiltrating lobular carcinoma-1 if histology was infiltrating lobular carcinoma; 0 otherwise.

PATH9-Inflammatory carcinoma-1 if histology was inflammatory carcinoma; 0 otherwise.

PATH10-Other carcinomas-1 if histology was any other single type of carcinoma; 0 otherwise.

DIFF-Differentiation-3 ordered categories for differentiation of the carcinoma.

FOCI-Foci of disease-1 if disease is unicentric; 0 if disease is multicentric.

CELL-Size of cells-1 if cancer cells are small cell; 0 if the cells are large cell.

INFIL-Infiltration-3 ordered categories for the amount of lymphocytic infiltration of the primary tumor.

These 80 variables were the input data for exploratory work with the discriminant analysis to reduce further the dimension of the problem. The results of the analysis appear in Chapter 5.

4.4 Missing Data

The problem of missing data was acute in this study. History taking by the physician who first examined the patient at BCCI was uneven in quality. Some histories were quite complete, while others had only a few main points covered. Patients also varied in how much they could remember or wished to tell. Some older patients (in their sixties and seventies) could remember little of what had happened to them. Any incident that could be verified by hospital or physician records was checked by the medical records department at BCCI. Because of the length of time since diagnosis (up to 22 years) and the fact that 68 percent of the patients were dead at the time of the study, no further followup was attempted for missing data.

When information about a variable was missing, a numerical missing-data code was used. After the data were collected, an examination of the variables with many missing data entries was undertaken. A variable that had not been mentioned in the chart had been coded as missing. In some cases the variable was not mentioned because the patient did not have that attribute. For example, if smoking was not mentioned in the history, it was much more likely that the patient did not smoke than that the patient smoked and the fact had not been reported. Six variables[3] were found to be of that type and so the missing-data code was changed to the code for absence of the attribute. Other variables that had missing data values were not changed. A patient with missing data for a particular variable was kept in the analysis when that variable was not included and

---

[3]The six variables that were changed in this manner were SMOKE, DYSMEN, OTHDRG, HORMON, MAZODY, and MASTIT.

was dropped from the analysis when the variable was included. More will be said about inclusion and exclusion of cases in Chapter 5.

4.5 Sources of Error

Four main sources of error existed for the data. As stated in the previous section, the patients and doctors were the two primary sources of error. The errors in both cases are biased toward omitting data. A patient was unlikely to claim a disease or operation that had not occurred and most of the claims were confirmed by followup or removed from the history. The doctors had no reason to claim things that did not occur. However, it was quite easy for the patient to forget an incident occurring many years prior to diagnosis and for the physician to neglect to ask specifically for all possibilities.

A third source of error was in the data coding. Since all the data were collected by the same individual, any systematic errors in interpretation should be consistent for all patients. If an error in interpretation of the medical facts occurred, it was the same in each instance. It is assumed that the other data coding errors were random.

The final main source of error in data collection was the keypunching of the data. To minimize such errors, all data were proofread after keypunching. A program was written that printed the numbers from the cards in the same format as the data coding form. The computer output was then compared to the original data coding forms to find errors. All errors found in the keypunching were corrected before the analysis was begun. Thus, there were few remaining errors because of the keypunching.

Chapter 5

DATA ANALYSIS

5.1 Computer Programs

For the theoretical and practical reasons discussed in Chapter 3, it was decided that stepwise discriminant analysis and logistic regression would be used for the classification of patients by nodal status. The linear discrimination was done using the Biomed Stepwise Discriminant Analysis program (BMD07M). It performed a two-group discriminant analysis wherein the variables were selected so as to maximally reduce the remaining variance. The classification functions which included those selected variables were then used to classify the cases.

Logistic regression was accomplished by the use of a log-linear model program developed by M. Nerlove and S.J. Press. A listing of that program appears in their report [35, p. 101 ff]. The coefficients of the logistic probability function were estimated by the Newton-Raphson algorithm, where the starting values were found by ordinary least squares. The program produced estimates of the logistic probability coefficients. A user-written program was then used to calculate the posterior probability of being in group 1 and thus the classification for each case.

5.2 Selecting Cases

In order to avoid using patients with missing data for the included variables, a procedure was devised for selecting cases. Instead of eliminating cases with missing values, group means could have been substituted for missing values. It was felt that that would not be appropriate here because of the nature of the data. The large number of binary variables caused problems for this approach. The mean of a 0-1 variable calculated for many patients would be between 0 and 1. A value between 0 and 1 for the binary variables was deemed unacceptable.

All cases with complete information for all variables were chosen for a preliminary analysis. There were 173 such cases. Discriminant analysis was run on these 173 cases to select the variables that were of some significance. Significance was defined quite liberally (F probability to enter $\leq .20$) so that even variables of marginal significance would be included initially. For reasons presented in Chapter 3 the subset of variables chosen in this manner should have included all the variables for which the coefficients were significantly different from zero. It will also include some for which the coefficient should have been estimated as zero. The variables which had coefficients that were not significant were eliminated in the final analysis by the logistic regression. The results of the discriminant analysis are presented in Table XI and XII. Fourteen variables were chosen by this process. One variable was entered in the third step and then removed in the twelveth step. It was decided to include that variable in further work in case it proved to be significant later.

Logistic regression was then done with the fifteen variables and 173 cases to compare to the discriminant analysis. The results are

Table XI

Variables Chosen by Linear Discrimination for 173 Cases

Variables	F Probability to Enter
1. THYROD	.0038
2. BRSIS	.0121
3. DURTON	.0114
4. NODEPL	.0199
5. HEART	.0640
6. LOC1	.1020
7. SYMPT2	.0586
8. SYMPT3	.0265
9. KIDNEY	.0743
10. BIRTHS	.1435
11. HORMON	.0820
12. MASTIT	.0970
13. BRMOM	.1087
14. OTHILL	.1746

Variable entered and then removed:

1. MONNUR

Table XII

Classification of 173 Cases by Linear Discrimination

		Classified to group 0	1	
Actual group	0	90	17	107
	1	33	33	66
		123	50	173

CPU time 11.292 sec.

Correct classification 71.10%

shown in Table XIII and XIV. Nine of the fifteen variables were chosen as significant ($P < .05$). The CPU time was twice as long as for linear discrimination. However, the classification by logistic regression was better than the classification by discriminant analysis — 77 percent correct for logistic regression and 71 percent correct for discriminant analysis. Both procedures did poorly in classifying those known to have positive nodes. Discriminant analysis classified only 50 percent of such cases correctly, while logistic regression classified 59 percent of them correctly. For those known to have negative nodes the correct classification rates were 84 percent and 89 percent, respectively.

The fifteen variables chosen by linear discrimination (Table XI) were then used for a new case-selection procedure. Patients who had complete information for those variables were chosen for further analysis. A total of 503 patients had complete information on these fifteen variables. This group of patients was the sample for the final classification procedure.

5.3 Classification of 503 Cases

Discriminant analysis was done for the 503 selected cases and fifteen variables. Again the level of significance was defined liberally — F probability to enter less than or equal to .20. The results appear in Tables XV and XVI. Only four variables were significant enough to enter the discriminant function. The discriminating power of the function was even worse for those with positive nodes; only 12 percent were correctly classified into group 1. For negative nodes there was 96 percent correct classification.

Table XIII

Variables Chosen by Logistic Regression for 173 Cases

Variables	Asymptotic Significance
1. DURTON	.03017
2. NODEPL	.03290
3. HEART	.01377
4. SYMPT3	.00338
5. KIDNEY	.04828
6. BIRTHS	.01239
7. HORMON	.08840
8. MASTIT	.01727
9. BRMOM	.12352

Table XIV

Classification of 173 Cases by

Logistic Regression

		Classified to group		
		0	1	
Actual	0	95	12	107
group	1	27	39	66
		122	51	173

CPU time	25.652 sec.
Correct classification	77.46%
Iterations	39

Table XV

Variables Chosen by Linear Discrimination for 503 Cases

Variables	F Probability to Enter
1. NODEPL	.0001
2. SYMPT3	.0007
3. THYROD	.0171
4. HEART	.0432

Table XVI

Classification of 503 Cases by Linear Discrimination

		Classified to group 0	1	
Actual group	0	321	12	333
	1	149	21	170
		470	33	503

CPU time 4.714 sec.

Correct classification 67.99%

Table XVII

Variables Chosen by Logistic Regression for 503 Cases

Variables	Asymptotic Significance
1. NODEPL	.00016
2. SYMPT3	.00086
3. THYROD	.02031
4. HEART	.04636

Table XVIII

Classification of 503 Cases by Logistic Regression with 15 Variables

		Classified to group 0	1	
Actual group	0	318	15	333
	1	147	23	170
		465	38	503

CPU time	18.251 sec.
Correct classification	67.78%
Iterations	10
Log of the likelihood	-301.774846

Logistic regression was run for the same fifteen variables and 503 cases. The results are shown in Tables XVII and XVIII. A comparison of Tables XV and XVII shows that the same four variables were significant and in the same order of significance. Again classification was much worse for positive nodes than negative nodes — 95 percent correct for negative nodes and 14 percent correct for positive nodes.

Since only four of the variables were significant, it was decided to rerun the logistic regression with only those four variables included. The previous run had forced all fifteen variables into the classification function regardless of significance. The results are given in Table XIX. All variables remained highly significant ($P < .05$). Using the

Table XIX

Classification of 503 Cases by Logistic Regression with with 4 Variables

		Classified to group 0	1	
Actual group	0	324	9	333
	1	150	20	170
		474	29	503

CPU time	9.810 sec.
Correct classification	68.39%
Iterations	8
Log of the likelihood	-303.560423

log-likelihood test to determine whether the ten β_i dropped from the second model should be zero, it was found from Tables XVIII and XIX that $U = -2\lambda = -2(-303.560423 + 301.774846) = 3.57114$.[1] The .05 level of significance chi-square value for ten degrees of freedom is 18.307. Thus, we conclude that the reduced model is a good fit when compared to the "full" model with fourteen variables. Correct classification for positive nodes was 12 percent and correct classification for negative nodes was 97 percent. A comparison of Tables XVI and XIX shows that logistic regression was marginally better than discriminant analysis at classification and took twice as long.

For both discriminant analysis and logistic regression there was a drop in discriminating power with the increase in number of cases. Linear discrimination went from 71.10 percent to 67.99 percent while logistic regression went from 77.46 percent to 68.39 percent. Part of the drop can be explained by the increase in the number of cases.[2] At least for logistic regression there appears to be some other factor influencing the classification. A likely contributing factor is less reliable data for the 330 added patients. The additional cases had some missing data values for variables not considered in the function and thus

---

[1] $\lambda = \log_e$ (likelihood).

[2] When doing a linear regression, one calculates an adjusted R^2 to account for differences in numbers of observations. In that case $R^2_{adj} = 1 - (1-R^2)\frac{N-1}{N-p}$ where N is the number of cases and p is the number of terms in the model [33]. For the numbers we were concerned with here the factor $\frac{N-1}{N-p}$ would be 1.0813 for 173 cases and 1.0060 for 503 cases. Thus, only small differences would be attributed to the increase in number of cases.

their information may be less reliable even when a variable was observed. That is, a patient who admits to not knowing certain information may be unreliable for other answers that were given. Also a doctor who looks for general answers may not ask for elaboration of answers to ensure complete recording of data.

5.4 Subsets of Patients for Further Classification

There is much medical evidence that breast cancer runs a much different course in premenopausal and postmenopausal women. "The age-specific incidence and mortality curves of breast cancer have two components. The premenopausal component is steeper. . . . The postmenopausal slope is less steep than the premenopausal. . ." [9, p. 721]. It has been hypothesized that the premenopausal carcinomas are hormone dependent, while the postmenopausal tumors are not. Evidence for the hypothesis comes from treatment results. Oophorectomy (removal of the ovaries and consequent cessation of estrogen production) is a successful treatment adjunct for premenopausal women. It seems to make no difference in postmenopausal women since the ovaries have already ceased production of estrogen.

In the light of the medical evidence, it was surprising that neither age nor menopausal status appeared as a significant variable. Apparently, this factor was confounded by the other factors in the model. Consequently, it was decided to investigate the menopausal status further. A stratification of patients by menopausal status was a possible avenue of investigation. The first step was to see if there were two distinct components of age incidence. A histogram of incidence versus age

(Figure 5.1) showed the possiblity of two components. Histograms were produced for the two subsets: premenopausal patients (Figure 5.2) and postmenopausal patients (Figure 5.3). Examination of Figure 5.2 and Figure 5.3 showed clear differences. The premenopausal patients had a graph with a very steep gradient while the postmenopausal patients had a graph that was more nearly flat for a twenty year period. The difference between the slopes may be because of different "censors" in the two groups — the premenopausal patients are censored by menopause while the postmenopausal patients are censored by death. However, other investigators [36,28] had found significant differences in the course of the disease between premenopausal and postmenopausal patients. Thus, it was decided to divide the patients on the basis of menopausal status and do the analyses on each subset of patients separately.

The 173 patients with full information were separated for preliminary work on the basis of menopausal status. Sixty of them were postmenopausal patients and 113 were premenopausal patients. Discriminant analysis was then done for the two subsets separately to reduce the dimension of the problem. Tables XX and XXI show the variables chosen for postmenopausal and premenopausal patients, respectively. Only five of the variables were chosen for both groups: DURTON, KIDNEY, BIRTHS, MASTIT, and HEART. Tables XXII and XXIII present the results of classification for the groups separately. The percentage classified correctly was greater for each group than when the classification was for the combined group (88 percent and 80 percent versus 71 percent for combined). Consequently, each group was investigated with a different set of variables.

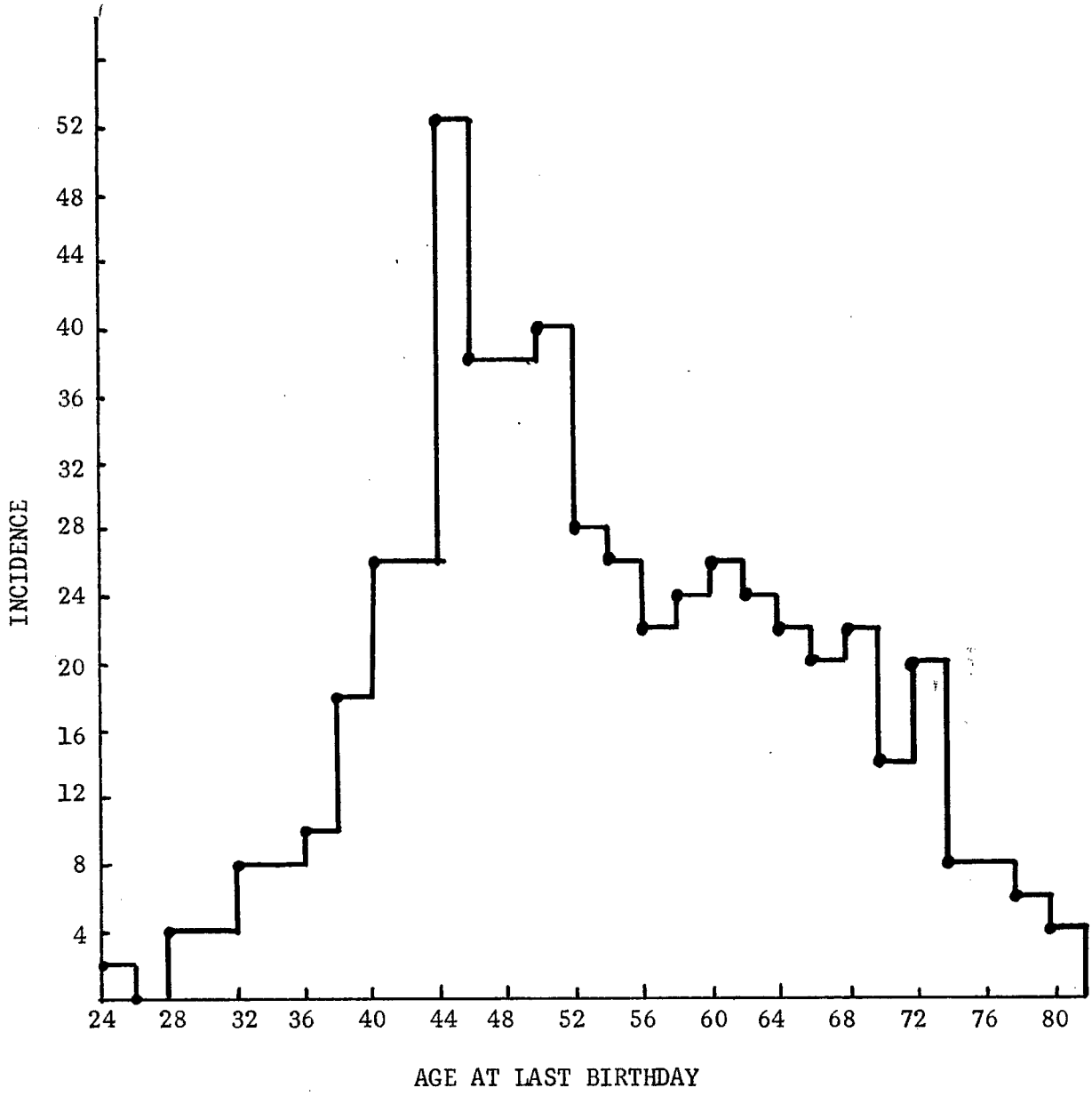


Figure 5.1. Histogram of age-incidence for all 535 cases.

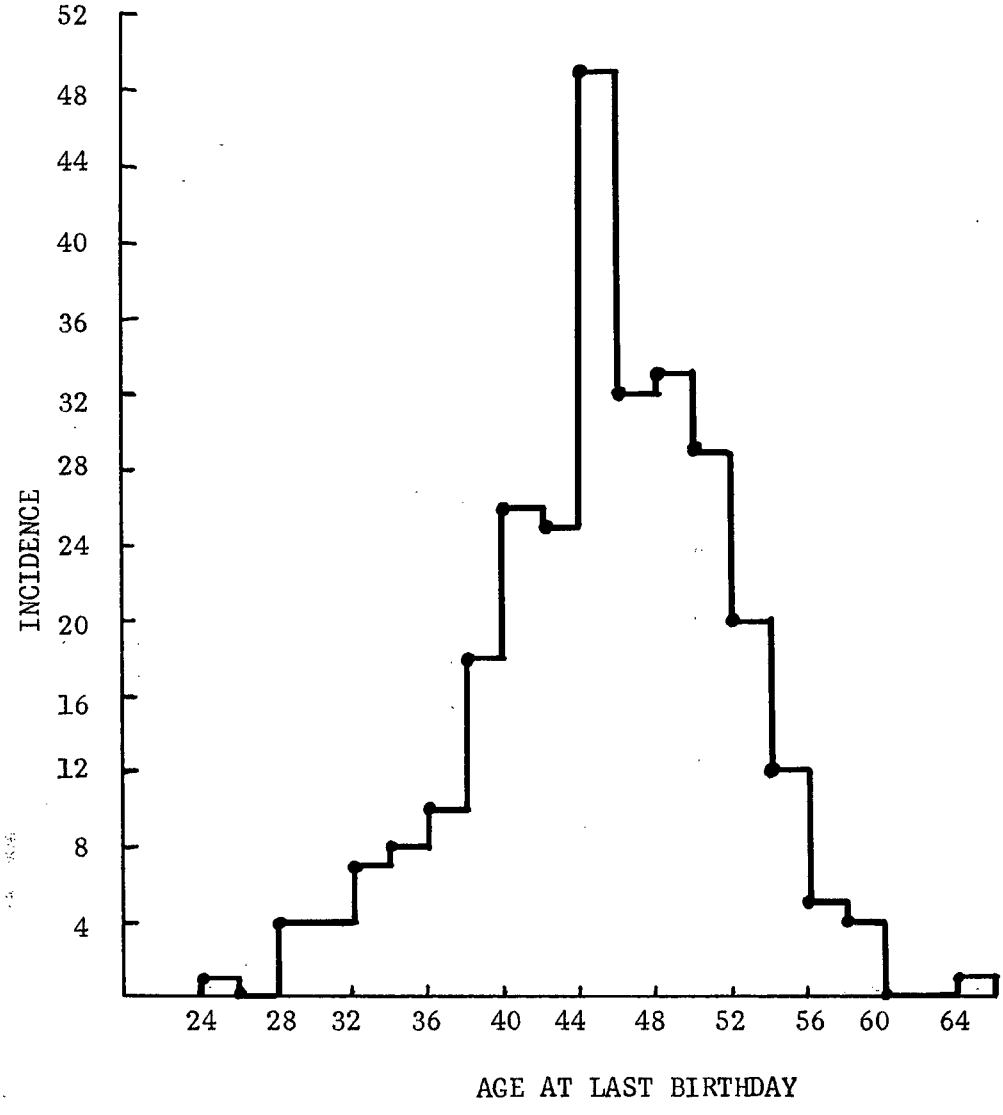


Figure 5.2. Histogram of age-incidence for premenopausal patients.

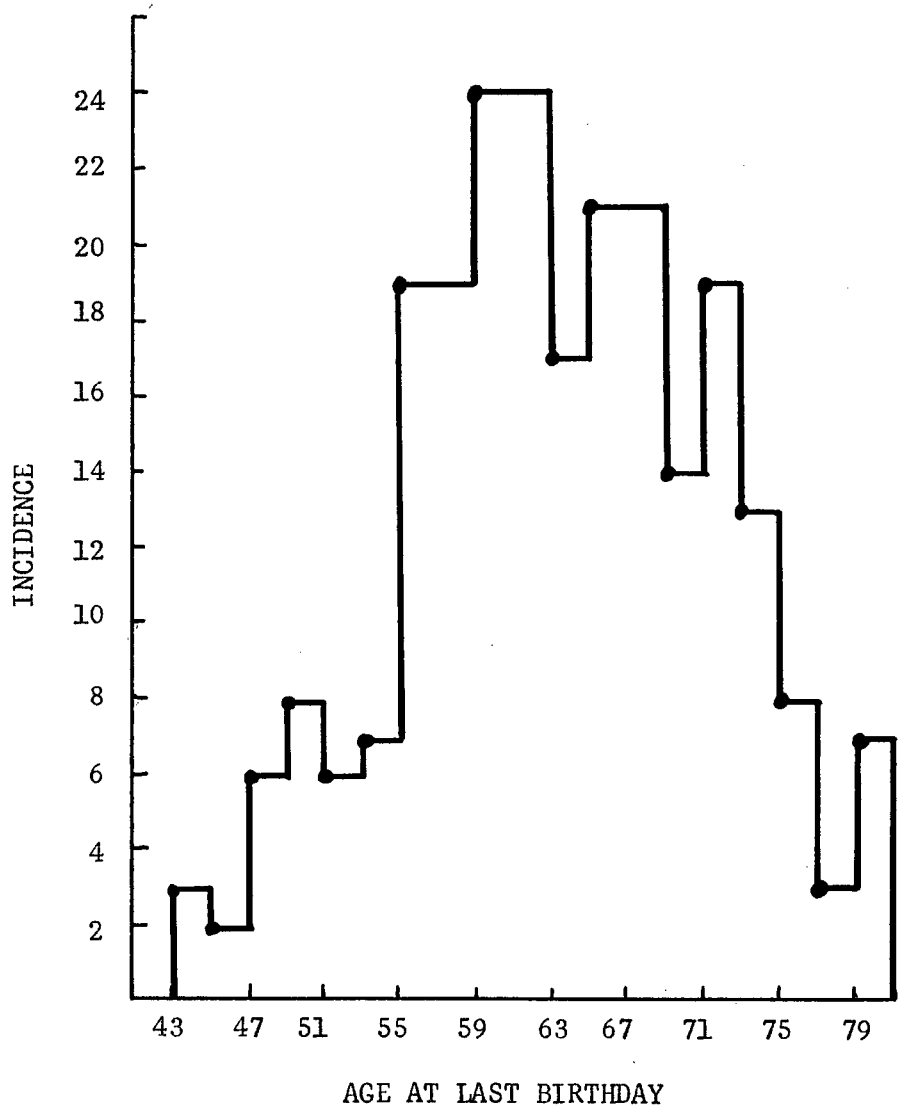


Figure 5.3. Histogram of age-incidence for postmenopausal patients.

Table XX

Variables Chosen by Discriminant Analysis for 60 Post-menopausal Patients

Variables	F Probability to Enter
1. HEART	.0155
2. NODEPL	.0054
3. BIRTHS	.0373
4. SKIN	.0529
5. SMOKE	.0758
6. TRAUMA	.0283
7. OTHSUR	.0824
8. BRFED	.0729
9. DURTON	.0287
10. BRSIS	.1019
11. LYMPH	.0471
12. OOPHOR	.0646
13. KIDNEY	.0745
14. AGE	.0738
15. BENIGN	.0481
16. SIZE	.0936
17. MASTIT	.1441
18. RACE	.0582
19. CELL	.1760

Variable entered and then removed:

1. CANCER

Table XXI

Variables Chosen by Discriminant Analysis for 113 Premenopausal Patients

Variables	F Probability to Enter
1. SYMPT3	.0013
2. THYROD	.0088
3. BREAST	.0398
4. DURTON	.0341
5. PATH6	.0538
6. PELVIC	.0539
7. KIDNEY	.0538
8. HORMON	.0326
9. BIRTHS	.0395
10. BROTH	.0713
11. BRMOM	.0647
12. MASTIT	.0125
13. PATH2	.0860
14. HEART	.0808
15. REG	.1089
16. ALLERG	.0744

Table XXII

Classification of 60 Postmenopausal Patients by Linear Discrimination with 19 Variables

		Classified to group 0	1	
Actual group	0	32	5	37
	1	2	21	23
		34	26	60

CPU time	7.14 sec.
Correct classification	88.33%

Table XXIII

Classification of 113 Premenopausal Patients by Linear Discrimination with 16 Variables

		Classified to group 0	1	
Actual group	0	61	9	70
	1	14	29	43
		75	38	113

CPU time	6.00 sec.
Correct classification	79.65%

5.5 Classification of Postmenopausal Patients

Nineteen variables were selected for the postmenopausal patients. Since the variables were chosen in the order of their explanatory power, it was decided to use only the first sixteen variables for the logistic regression. The discrimination programs could have been run twice on subsets of the nineteen variables to pick the best sixteen, but previous work had indicated that that was not necessary. All postmenopausal patients were screened for full information on the nineteen variables, yielding a total of 128 cases.

Discriminant analysis was run on the 128 patients with nineteen variables. Four variables were chosen for the analysis and 72.66 percent were correctly classified (Tables XXIV and XXV). Again negative nodes were classified better (80 percent correct) than positive nodes were (60 percent correct). Then logistic regression was performed with sixteen variables. The results appear in Tables XXVI and XXVII. Only three variables were found to be significant. Logistic regression was run again with the three significant variables (Table XXVIII). A comparison of Tables XXV and XXVIII shows that logistic regression was poorer at classifying than discriminant analysis was. Since different variables had been used it was decided to rerun logistic regression with the four variables of Table XXIV. The results of that run are in Table XXIX. A comparison of Tables XXV and XXIX shows that both methods classified the patients exactly the same.

To test whether certain β_i were zero, the log-likelihood test was used. From Tables XXVII and XXVIII the statistic for testing the reduction of the model to three variables was found to be $U = -2\lambda =$

Table XXIV

Subset of Variables Chosen by Discriminant Analysis for 128 Postmenopausal Patients

Variables	F Probability to Enter
1. NODEPL	.0001
2. AGE	.0638
3. BRSIS	.0682
4. SMOKE	.1934

Table XXV

Classification of 128 Postmenopausal Patients by Discriminant Analysis with 4 Variables

		Classified to group 0	1	
Actual group	0	65	16	81
	1	19	28	47
		84	44	128

CPU time	2.5 sec.
Correct classification	72.66%

Table XXVI

Subset of Variables Chosen by Logistic Regression for 128 Postmenopausal Patients

Variables	Asymptotic Significance
1. AGE	.05321
2. NODEPL	.00011
3. TRAUMA	.12377

Table XXVII

Classification of 128 Postmenopausal Patients by Logistic Regression with 16 Variables

		Classified to group 0	1	
Actual	0	68	13	81
group	1	19	28	47
		87	41	128

CPU time	32 sec.
Correct classification	75.00%
Iterations	37
Log of the likelihood	-67.1124784

Table XXVIII

Classification of 128 Postmenopausal Patients by Logistic Regression with 3 Variables

		Classified to group 0	1	
Actual group	0	64	17	81
	1	21	26	47
		85	43	128

CPU time	3.077 sec.
Correct classification	70.31%
Iterations	7
Log of the likelihood	-74.3288522

Table XXIX

Classification of 128 Postmenopausal Patients by Logistic Regression with 4 Variables

		Classified to group 0	1	
Actual	0	65	16	81
group	1	19	28	47
		84	44	128

CPU time	10.531 sec.
Correct classification	72.66%
Iterations	35
Log of the likelihood	-70.2542850

$-2(-74.3288522 + 67.1124784) = 14.432748$. The corresponding chi-square for twelve degrees of freedom and .05 significance level is 21.026. Thus, I concluded that the twelve β_i that were estimated as zero were zero. Using the same test and Tables XXVII and XXIX, the statistic is $U = -2(-70.2542850 + 67.1124784) = 6.283614$. Since $\chi^2_{.95}(11) = 19.675$, it was concluded that the reduced model with four variables was also acceptable.

5.6 Classification of Premenopausal Patients

A similar procedure was used for the premenopausal patients. The sixteen variables of Table XXI chosen by discriminant analysis were used to select full information cases. A total of 253 premenopausal patients had full information on those variables. When discriminant analysis was run with the 253 selected cases, ten variables were found to be significant. The results of the discrimination are shown in Tables XXX and XXXI. Table XXX shows the ten variables chosen as significant. From Table XXXI we see that again patients with negative nodes were classified better (95 percent correct) than patients with positive nodes (27 percent correct).

Logistic regression was then run with the 253 full information premenopausal patients and the same sixteen variables from Table XXI. The results of the logistic regression appear in Tables XXXII and XXXIII. Ten variables were significant in the logistic regression also. A comparison of Tables XXX and XXXII shows that nine of the variables are the same for both cases. The other variable in the logistic regression, DURTON, was the least significant of the ten variables. As with linear discrimination, classification of patients with positive nodes was poorer

Table XXX

Subset of Variables Chosen by Discriminant Analysis

for 253 Premenopausal Patients

Variables	F Probability to Enter
1. THYROD	.0013
2. SYMPT3	.0033
3. MASTIT	.0165
4. PELVIC	.0521
5. BRMOM	.0448
6. PATH2	.1186
7. KIDNEY	.1201
8. BIRTHS	.1012
9. HEART	.1132
10. BROTH	.1979

Table XXXI

Classification of 253 Premenopausal Patients by Discriminant Analysis with 10 Variables

		Classified to group 0	1	
Actual group	0	159	8	167
	1	63	23	86
		222	31	253

CPU time	2.5 sec.
Correct classification	71.94%

Table XXXII

Subset of Variables Chosen by Logistic Regression

for 253 Premenopausal Patients

Variables	Asymptotic Significance
1. BRMOM	.02721
2. BROTH	.13538
3. BIRTHS	.08456
4. HEART	.08576
5. KIDNEY	.11729
6. THYROD	.00284
7. PELVIC	.01061
8. MASTIT	.00059
9. DURTON	.19050
10. SYMPT3	.00726

Table XXXIII

Classification of 253 Premenopausal Patients by

Logistic Regression with 16 Variables

		Classified to group		
		0	1	
Actual	0	156	11	167
group	1	58	28	86
		214	39	253

CPU time	35.603 sec.
Correct classification	72.73%
Iterations	40
Log of the likelihood	-137.340218

than classification of patients with negative nodes (33 percent versus 93 percent).

In judging how well our models have worked for classification there are two types of tests. Goodness of fit tests are used to test whether certain β_i are zero. They test how well the model has been estimated. How well the estimated model classified is the second measure of the classification scheme. Although one gets a better fit of the model with more variables, the classification may be better with fewer variables in the model.

Since the previous logistic classification was with sixteen variables and only ten were significant, it was decided to rerun the logstic regression with only the ten significant variables. It was expected that that would classify better and the results presented in Table XXXIV confirmed that expectation. Logistic regression with ten variables classified better than logistic regression with sixteen variables or discriminant analysis with ten variables (74 percent versus 73 percent and 72 percent). To test whether the ten variable logistic model was a good enough fit compared to the sixteen variable logistic model, the log-likelihood test was used again. It was found from Tables XXXIII and XXXIV that $U = -2\ \lambda = -2(-140.582389 + 137.340218) = 6.48434$. The .05 significance level chi-square value for six degrees of freedom is 12.6. Thus, we can conclude that the reduced model provides a good fit.

5.7 Summary of Results

A summary of the classification results was prepared and appears in Table XXXV. The data analysis has confirmed the theoretical results

Table XXXIV

Classification of 253 Premenopausal Patients by Logistic Regression with 10 Variables

		Classified to group 0	1	
Actual group	0	162	5	167
	1	61	25	86
		223	30	253

CPU time	6.908 sec.
Correct classification	73.91%
Iterations	9
Log of the likelihood	-140.582389

Table XXXV

Summary of Classification Results

Model	Variables	# Cases	% Correct Positive Nodes	% Correct Negative Nodes	% Correct Overall
Combined premenopausal and postmenopausal:					
DA	14	173	50.00	84.11	71.10
LR	9	173	59.09	88.79	77.46
DA	4	503	12.35	96.40	67.99
LR	15	503	13.53	95.50	67.78
LR	4	503	11.76	97.30	68.39
Postmenopausal:					
DA	19	60	91.30	86.49	88.33
DA	4	128	59.57	80.25	72.66
LR	16	128	59.57	83.95	75.00
LR	3	128	55.32	79.01	70.31
LR	4	128	59.57	80.25	72.66
Premenopausal:					
DA	16	113	67.44	87.14	79.65
DA	10	253	26.74	95.21	71.94
LR	16	253	32.56	93.41	72.73
LR	10	253	29.07	97.01	73.91

DA = Discriminant Analysis

LR = Logistic Regression

of Chapter 3. Linear discrimination was an efficient method for the preliminary analyses. For the final analyses logistic regression provided more correct classifications with a significant time increase (in one case the classification was the same). However, classification in all cases was less than hoped for. Examination of the cases misclassified showed that many of such cases were near the boundary. That suggested the use of a two-stage procedure as described in Chapter 3. Patients with high or low posterior probabilities could be classified on the basis of this data. The patients with probabilities near .05 would need further data before classification could be done.

Tables XXXVI and XXXVII were prepared to demonstrate some possible two-stage procedures. Table XXXVI used the logistic classification for 253 premenopausal patients with ten variables. Table XXXVII used the logistic classification for 128 postmenopausal patients with four variables. For each decile of posterior probability the numbers of patients correctly and incorrectly classified were tabulated. If an error rate of ten percent is acceptable, then the boundaries could be set at .4 and .7 for Table XXXVI. That is, all patients with posterior probability less than .4 would be classified as having negative nodes. All patients with posterior probability greater than .7 would be classified as having positive nodes. The patients in the middle would need further investigation before classification. In this case, that would be 109 patients or 43 percent. A similar analysis of Table XXXVII produced boundary points of .3 and .7 for a nine percent error rate and 73 patients (57 percent) to have further investigation.

Table XXXVI

Number of Premenopausal Patients Correctly and Incorrectly Classified by Decile of Posterior Probability

Decile	Correct	Incorrect
.00 - .1	14	0
.11 - .2	31	8
.21 - .3	51	14
.31 - .4	42	21
.41 - .5	21	17
.51 - .6	6	2
.61 - .7	3	0
.71 - .8	11	0
.81 - .9	5	2
.91 - 1.0	4	1

Table XXXVII

Number of Postmenopausal Patients Correctly and Incorrectly Classified by Decile of Posterior Probability

Decile	Correct	Incorrect
.00 - .1	10	0
.11 - .2	18	5
.21 - .3	19	4
.31 - .4	10	7
.41 - .5	8	4
.51 - .6	12	9
.61 - .7	14	5
.71 - .8	1	1
.81 - .9	0	1
.91 - 1.0	0	0

Examination of the cases that were misclassified also showed runs of errors. For one group of patients 90 percent of the errors occurred in two consecutive years of the nine years studied and 95 percent of the errors were in three consecutive years of the nine years. That lends credence to the idea that certain history takers were better than others. It is probable that better classification could have been achieved with better data for those years.

The medical implications of the analysis will be discussed in the next chapter.

Chapter 6

CONCLUSIONS

For the medical classification problem of distinguishing between those breast cancer patients with supraclavicular or internal mammary lymph node metastases and those without such metastases, two models were selected — linear discrimination and logistic regression. The empirical results verified the theoretical findings that linear discrimination was faster than logistic regression but logistic regression provided a greater proportion of correct classifications.

When the classification was done on 503 cases with full information, approximately 68 percent correct classification was achieved. While this is better than the clincial staging, it was not as good as had been hoped for. Dividing the patients by menopausal status and classifying the groups separately provided better classification because of the differing disease processes in the two groups. For 253 premenopausal patients the proportion of correct classifications was 74 percent. For 128 postmenopausal patients the correct proportion was 75 percent. Thus, separating the groups on the basis of menopausal status provided classification that was better than when the groups were combined.

A two-stage procedure was proposed to make the error rate smaller. Patients were classified on the basis of the data into three groups: those with negative nodes, those with positive nodes, and those for whom more data had to be collected before a final classification was made. With an error rate of 10 percent or less, 43 percent of the premenopausal and 57 percent of the postmenopausal patients required further observation.

It was concluded that for a medical diagnosis problem such as this with data that are clearly non-normal and often not even continuous, the use of linear discriminant analysis was adequate for the exploratory work and to reduce the dimension of the problem. In the preliminary analyses of the subgroups of premenopausal and postmenopausal patients with small numbers of patients and complete information, discriminant analysis provided more correct classifications for positive nodes. Logistic regression was preferred for the final analyses since it provided better estimators and consequently more classifications that were correct when there were more patients and less complete information about the patients.

The medical conclusions are not so definitive. The classification procedures used here suggest areas where further investigation would be useful. For all patients combined the variables that entered the final analysis were NODEPL, SYMPT3, THYROD, and HEART. As expected palpable lymph nodes were positively correlated with pathologically involved lymph nodes. The presence of changes in the nipple or discharge from the nipple as the first symptom was positively correlated with positive nodes. A history of thyroid disease was also positively correlated with positive nodes. A history of heart disease was negatively correlated

with positive nodes. Thus, the physician might consider these factors when trying to evaluate the nodes clinically.

When the premenopausal and postmenopausal patients were considered separately, some other factors were suggested. For the premenopausal patients the variables that entered the final analyses were THYROD, SYMPT3, MASTIT, PELVIC, BRMOM, PATH2, KIDNEY, BIRTHS, HEART, and BROTH. Again a history of thyroid disease was positively correlated with positive nodes and nipple change or discharge as the first symptom was positively correlated with positive nodes. A history of benign breast disease during lactation was positively correlated with involved nodes. Previous pelvic surgery was negatively correlated with involved nodes. Breast cancer in the patient's mother and other relatives were both negatively correlated with involved nodes. This was probably a result of the patient's increased awareness of the disease and consequent earlier diagnosis. Noninfiltrating papillary carcinoma was highly negatively correlated with positive nodes. That is, patients with that pathological type of carcinoma rarely had nodal metastases. A history of kidney disease was negatively correlated with positive nodes. The number of full term pregnancies was slightly negatively correlated with positive nodes. Heart disease was again negatively correlated with involved nodes.

For the post-menopausal patients the variables that entered the final analyses were NODEPL, AGE, BRSIS, and SMOKE. Again as expected nodes palpable was positively correlated with positive nodes. Age was very slightly negatively correlated with positive nodes. Breast cancer in the patient's sister was negatively correlated with positive nodes. Smoking was slightly negatively correlated with positive nodes. The

disease in the postmenopausal patients was less variable than the disease in the premenopausal patients. The lesser degree of variability resulted in more accurate classifications for the postmenopausal patients.

The results presented above could be used to help the physician in his clinical diagnosis of the nodal status. They also suggest areas in which further research could be done.

BIBLIOGRAPHY

[1] Anderson, J.A., "Separate Sample Logistic Discrimination," *Biometrika*, 59 (1972), 19-35.

[2] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Inc., New York, 1958.

[3] Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Mass., 1975.

[4] Bock, R.D., "Estimating Multinomial Response Relations," *Essays in Probability and Statistics*, ed. by R.C. Bose, *et al.*, University of North Carolina Press, Chapel Hill, North Carolina, 1970.

[5] Brinkley, D. and J.L. Haybittle, "A 15-year Follow-up Study of Patients Treated for Carcinoma of the Breast," *British Journal of Radiology*, 41 (1968), 215-221.

[6] Brinkley, D. and J.L. Haybittle, "The Curability of Breast Cancer," *The Lancet*, July 19, 1975, 95-97.

[7] Cancer Research Campaign, "Management of Early Cancer of the Breast," *British Medical Journal*, 1 (1976), 1035-1038.

[8] Cochran, W.G. and C.E. Hopkins, "Some Classification Methods with Multivariate Qualitative Data," *Biometrics*, 17 (1961), 10-32.

[9] Correa, P., "The Epidemiology of Cancer of the Breast," *American Journal of Clinical Pathology*, 64 (1975), 720-727.

[10] Cox, D.R., *Analysis of Binary Data*, Methuen and Co., Ltd., London, 1970.

[11] Crawford, G., "Breast Cancer: Selective Biopsy — Rationale and Results," Cancer Control Agency Seminar, 1976.

[12] Efron, B., "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70 (1975), 892-898.

[13] Federer, W.T., "Procedures and Designs for Screening Material in Selection and Allocation with a Bibliography," *Biometrics*, 19 (1963), 553-587.

[14] Finney, D.J., R. Latscha, B.M. Bennett, and P. Hsu, *Tables for Testing Significance in a 2x2 Contingency Table*, Cambridge University Press, New York, 1963.

[15] Fisher, E.R., R.M. Gregorio, and B. Fisher, "The Pathology of Invasive Breast Cancer — A Syllabus Derived from Findings of the National Surgical Adjuvant Breast Project (Protocol Number 4)," *Cancer*, 36 (1975), 1-85.

[16] Fisher, E.R., R.M. Gregorio, C. Redmond, W.S. Kim, and B. Fisher, "The Significance of Extranodal Extension of Axillary Metastases," *American Journal of Clinical Pathology*, 65 (1976), 439-444.

[17] Friel, J.P. (ed.), *Dorland's Illustrated Medical Dictionary*, Twenty-fifth Edition, W.B. Saunders Co., Toronto, 1974.

[18] Gardner, M.J. and D.J.P. Barker, "Diagnosis of Hypothyroidism: A Comparison of Statistical Techniques," *British Medical Journal*, 2 (1975), 260-262.

[19] Goldstein, M. and M. Rabinowitz, "Selection of Variates for the Two-Group Multinomial Classification Problem," *Journal of the American Statistical Association*, 70 (1975), 776-781.

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

[21] Guzeman, L.F., B.C. Peters, and H.F. Walker, "On Minimizing the Probability of Misclassification for Linear Feature Selection," *The Annals of Statistics*, 3 (1975), 661-668.

[22] Haagensen, D.C., "The Choice of Treatment for Operable Carcinoma of the Breast," *Surgery*, 76 (1974), 685-714.

[23] Haagensen, D.C., E. Cooley, *et al.*, "Treatment of Early Mammary Carcinoma: A Cooperative International Study," *Annals of Surgery*, 170 (1969), 875-890.

[24] Haberman, S.J., *The Analysis of Frequency Data*, University of Chicago Press, Chicago, 1974.

[25] Halperin, M., W.C. Blackwelder, and J.I. Verter, "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches," *Journal of Chronic Diseases*, 24 (1971), 125-158.

[26] Hartz, S.C., and L.A. Rosenberg, "Computation of MLE for the Multiple Logistic Risk Function for Use with Categorical Data," *Journal of Chronic Diseases*, 28 (1975), 421-430.

[27] Krzanowski, W.J., "Discrimination and Classification Using Both Binary and Continuous Variables," *Journal of the American Statistical Association*, 70 (1975), 782-790.

[28] Lewison, E.F., *Breast Cancer and its Diagnosis and Treatment*, The Williams and Wilkins Company, Baltimore, 1955.

[29] McCabe, G.P., "Computations for Variable Selection in Discriminant Analysis," *Technometrics*, 17 (1975), 103-109.

[30] McCabe, G.P. and R.J. Pohl, *A Computer Program for Variable Selection in Discriminant Analysis*, Purdue University Department of Statistics Mimeograph Series, Number 334, Purdue University, Lafayette, Indiana, 1973.

[31] McDivitt, R.W., F.W. Stewart, and J.W. Berg, *Tumors of the Breast*, Armed Forces Institute of Pathology, Bethesda, Maryland, 1968.

[32] Maehle, B.O. and F. Harviet, "Prognostic Typing in Breast Cancer," *Journal of Clinical Pathology*, 26 (1973), 784-791.

[33] Montgomery, D.B. and D.G. Morrison, "A Note on Adjusting R^2," *Journal of Finance*, 1973, 1009-1013.

[34] Mueller, C. and W. Jeffries, "Cancer of the Breast: Its Outcome as Measured by the Rate of Dying and Causes of Death," *Annals of Surgery*, 182 (1975), 334-341.

[35] Nerlove, M. and S.J. Press, *Univariate and Multivariate Log-Linear and Logistic Models*, RAND Report R-1306-EDA/NIH, The RAND Corporation, Santa Monica, California, 1973.

[36] Papaioannou, A.N., *The Etiology of Human Breast Cancer*, Springer-Verlag, New York, 1974.

[37] Peters, M.V., "Cutting the 'Gordian Knot' in Early Breast Cancer," *Annals of the Royal College of Physicians and Surgions of Canada*, 8 (1975), 186-192.

[38] Poser, C.M., *et al.*, "Amino Acid Residues of Serum and CSF Protein in Multiple Sclerosis: Clinical Application of Statistical Discriminant Analysis," *Archives of Neurology*, 32 (1975), 308-314.

[39] Prentice, R., "Use of the Logistic Model in Retrospective Studies," unpublished manuscript, University of Washington.

[40] Press, S.J., *Applied Multivariate Analysis*, Holt, Rinehart and Winston, New York, 1972.

[41] Press, S.J. and S. Wilson, "Choosing Between Logistic Regression and Discriminant Analysis," to be published, manuscript at The University of British Columbia, 1977.

[42] Ramberg, J.S. and J.D. Broffitt, "Selecting the Best Set of Linear Discriminant Variates," *Proceedings of Computer Science and Statistics: 8th Symposium on the Interface*, 257-261.

[43] Smith, D.C., R. Prentice, D.J. Thompson, W.L. Herrmann, "Association of Exogenous Estrogen and Endometrial Carcinoma," *The New England Journal of Medicine*, 293, 1164-1167.

[44] Snedecor, G.W. and W.G. Cochran, *Statistical Methods*, Sixth Edition, The Iowa State University Press, Ames, Iowa, 1967.

[45] Tallis, G.M., P. Leppard, and G. Sarfaty, "A General Classification Model with Specific Application to Response to Adrenalectomy in Women with Breast Cancer," *Computers and Biomedical Research*, 8 (1975), 1-7.

[46] Tallis, G.M. and G. Sarfaty, "On the Distribution of the Time to Reporting Cancers with Application to Breast Cancer in Women," *Mathematical Biosciences*, 19 (1974), 371-376.

[47] Truett, J., J. Cornfield, and W.B. Kannel, "A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham," *Journal of Chronic Diseases*, 20 (1967), 511-524.

[48] VanNess, J.W. and C. Simpson, "On the Effects of Dimension in Discriminant Analysis," *Technometrics*, 18 (1976), 175-187.

[49] Walker, S.H. and D.B. Duncan, "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika*, 54 (1967), 167-179.

[50] Ziel, H.K. and W.D. Finkle, "Increased Risk of Endometrial Carcinoma Among Users of Conjugated Estrogens," *The New England Journal of Medicine*, 293 (1975), 1167-1170.

[51] Zielezny, M. and O.J. Dunn, "Cost Evaluation of a Two-Stage Classification Procedure," *Biometrics*, 31 (1975), 37-47.

APPENDIX A

TREATMENT STUDY GROUPS

		Number of cases
A.	No mastectomy — Standard radiation	148
B.	Simple mastectomy — Standard radiation	13
C.	Radical mastectomy — Standard radiation	316
D.	Radical mastectomy — Radiation which did not include axilla	17
E.	Radical mastectomy — with preoperative radiation	4
F.	Extended radical mastectomy — Radiation to supraclavicular only	1
G.	Simple mastectomy — No radiation	1
H.	Radical mastectomy — No radiation	20
I.	Simple mastectomy — with radiation, did not include chest wall	12
J.	Simple mastectomy — with radiation, did not include chest wall or axilla	1
K.	Simple mastectomy — Radiation and chemotherapy	1
L.	Hormones only	1
	TOTAL	535

APPENDIX B

MANCHESTER STAGING OF BREAST CANCER

Clinical

I Primary freely movable on contracted pectoral muscle or chest wall. Skin involvement, including ulceration, may be present but must be in direct continuity with the tumor and no extension wide of the tumor itself.

II As Stage I but there are palpable mobile lymph nodes in the axilla on the same side less than 2.5 cm.

III *Either* a) the skin invaded or fixed over an area wide of the tumor itself but still limited to the breast,
or b) the tumor fixed to underlying muscle but not to chest wall.
Axillary nodes, if present, must be mobile.

IV The growth has extended beyond the breast area as shown by:
a) Axillary nodes not mobile or >2.5 cm.
b) Tumor fixed to chest wall.
c) Supraclavicular node involvement.
d) Involvement of skin wide of breast.

e) Opposite breast involved with metastatic disease.

f) Distant metastases.

g) Inflammatory carcinoma.

Pagent's disease of nipple only is Stage I unless nodes present.

Pathological

I Disease confined to the breast.

II As in Stage I, plus metastatic disease confined to axillary lymph nodes below the level of the apex.

II? Level of axillary involvement unknown.

III Direct local spread from primary to:

a) skin wide of tumor.

b) underlying fascia or muscle.

IV

a) Direct extension from breast primary to rib or cartilage of chest wall.

b) Extension of disease beyond capsule of an axillary lymph node.

c) Involvement of apical or internal mammary lymph node or tissues.

d) Involvement of an axillary lymph node at any level which is found pathologically to be 2.5 cm. in size or large.

e) Distant metastases (including supraclavicular lymph nodes).

APPENDIX C

CODING INSTRUCTIONS AND DATA CODING FORM

Coding Instructions:

Occupation	
housewife	1
retired	2
technical & professional	3
clerical	4
laborer, outside	5
laborer, inside	6
other	7
unknown	9
Racial origin	
Caucasian	1
Negro	2
Indian	3
Asian	4
Semitic	5
Other	6
Unknown	9
Family history	
Yes	1
No	0
Unknown	9
Menopausal state	
Premenopausal & up to 5 years after	1
Postmenopausal 5 years	0
Age at menopause	
Premenopausal	88
Postmenopausal	Actual Years
Illnesses & surgery	
Yes	1
No	0
Unknown	9
First observation of symptom	
Patient	1
Medical professional	0
Duration of symptom	
1-97 months	Actual
≥ 98 months	98
Unknown	99
Tumor size	
< 2 cm	1
2 to 5 cm	2
> 5 cm	3
No lump palpable	4
Size not stated	9
Position of tumor	
Lower inner	1
Lower outer	2
Upper inner	3
Upper outer	4
Lymph node or tail	5
Nipple	6
Whole breast	7
Other	8
Unknown	9
Nodes, skin, & trauma	
Yes	1
No	0
Unknown	9
Breast	
Right	1
Left	0
Bilateral	3

First symptom
- Thickening 1
- Lump 2
- Pain 3
- Discharge from nipple 4
- Nipple inverted 5
- Skin changes 6
- Change in breast size 7
- Mammography, etc. 8
- Other 9

Overall body size
- Obese 1
- Average 2
- Slender 3

Breast size/shape
- Pendulous, very large 1
- Large, full 2
- Average 3
- Small 4

General physical condition
- Good 1
- Fair 2
- Poor 3

Nodal involvement
- Positive apical or i.m. nodes 1
- Positive lower axillary nodes 2
- No nodal involvement 3

Histological differentiation
- Well-differentiated 1
- Moderately differentiated 2
- Poorly- or undifferentiated 3
- Unknown 9

Foci of disease
- Unicentric 1
- Multicentric 0
- Unknown 9

Cell size
- Small 1
- Large 0
- Unknown 9

Cause of death
- Alive 0
- Breast cancer 1
- Intercurrent disease 2
- Lost to followup 3

Histology type

Paget's disease	1
Noninfiltrating papillary carcinoma	2
Infiltrating papillary carcinoma	3
Infiltrating duct carcinoma (scirrhus with productive fibrosis)	4
Adenocarcinoma	5
Colloid carcinoma (mucoid)	6
Medullary carcinoma	7
In situ lobular carcinoma	8
Infiltrating lobular carcinoma	9
Inflammatory carcinoma	10
Carcinoma, not otherwise specified	11
Other	12
Combinations of the above	13

Lymphocyte infiltrations in tumor

None	1
Minimal	2
Moderate or numerous	3
Unknown	9

CARCINOMA OF THE BREAST
SELECTIVE BIOPSY
1955 - 1963

CCARC Number
Card Identity
Anniversary Date
Date of Birth
Age at Anniversary
Clinical stage
Pathological stage
Marital status S-1,M-2,W-3,D-4
Occupation
Racial Origin
Number of Years in North America
Number of Years in B.C.

FAMILY HISTORY
Cancer other than breast
Father
Mother
Sister
Brother
Son
Daughter
Mat. Rel.
Pat. Rel.

Breast Cancer
Mother
Sister
Daughter
Other

Other diseases in family members
Diabetes
Tuberculosis
Heart Disease
Other.....

PERSONAL HISTORY
Smoker

Menstrual history
Menarche
Periods
Regular
Length

Interval between
Dysmenorrhoea
Drugs Hormones
Other

Menopause
Status pre 1 post 2
Age at menopause

Pregnancies
Age at first
Live births
Miscarriages
Number nursed
Months nursed

Patient breastfed

Serious illnesses
Diabetes
Tuberculosis
Heart disease
Typhoid
Hypertension
Ulcer
Kidney
Anemia
Asthma
Pneumonia
Childhood
Other

Surgery(not breast)
Oophorectomy
Tonsils
Appendix
Hysterectomy
Other pelvic
Cholecystectomy
Thyroidectomy
Adrenalectomy
Other.....

Previous breast ailment
Mastodynia during period
Mastitis in lactation
Benign breast disease not during lactation

History of Present Illness
First symptom
First observation of symptom
Duration of symptom
Tumour size-clinical
Position of tumour
Nodes palpable
Skin involvement
Breast
Trauma to breast

Patients Present Condition
Overall body size
Breast size/shape
General condition
Other illnesses present

PATHOLOGY
Nodal Involvement
Histology
Type
Differentiation
Foci of disease
Cell size

SURVIVAL
Date of death
Cause of death
Lost to followup

APPENDIX D

VARIABLES NOT INCLUDED IN THE ANALYSIS

General:

Year of diagnosis — year of initial diagnosis or treatment (anniversary)
Date of birth — month and year of birth (were used to check reported age but do not appear explicitly in functions)
Clinical stage — clinical stage (the factors used in staging appear as variables)
Pathological stage — pathological stage
Years in N.A. — number of years that patients has lived in North America
Years in B.C. — number of years that patient has lived in British Columbia

Family History:

Cancer other than breast — for father, mother, sister, brother, son, daughter, maternal relative, and paternal relative there was an indicator variable for occurrence of cancer and a variable for types of cancer
Other diseases in family members — diabetes, tuberculosis, heart disease, and other for blood relatives.

Menstrual history:

Menarche — age at which menstruation began
Lenth of periods — in days
Interval between — days between periods

Illnesses and Surgery of patient

Childhood diseases — mumps, measles, etc.
Typhoid
Ulcer — stomach ulcer
Tonsillectomy
Appendectomy

History of present illnesses:

Other symptoms — those appearing after the first
First observation of symptom — whether patient or medical professional first observed symptom of disease

Survival data: used to increase knowledge of history of problem but not pertinent to the classification problem

Date of Death
Cause of Death
Date lost to followup