STATISTICAL METHODS FOR DIAGNOSTIC TESTING: AN ILLUSTRATION USING A NEW METHOD FOR CANCER DETECTION

by

XIN SUN

PhD, Kansas State University, 2012

A THESIS

Submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2013

Approved by:

Major Professor
Gary Gadbury

Abstract

This report illustrates how to use two statistic methods to investigate the performance of a new technique to detect breast cancer and lung cancer at early stages. The two methods include logistic regression and classification and regression tree (CART). It is found that the technique is effective in detecting breast cancer and lung cancer, with both sensitivity and specificity close to 0.9. But the ability of this technique to predict the actual stages of cancer is low. The age variable improves the ability of logistic regression in predicting the existence of breast cancer for the samples used in this report. But since the sample sizes are small, it is impossible to conclude that including the age variable helps the prediction of breast cancer. Including the age variable does not improve the ability to predict the existence of lung cancer. If the age variable is excluded, CART and logistic regression give a very close result.

Table of Contents

List of Figures

List of Tables

Chapter 1 - Introduction to the project

1.1 Aim of the project

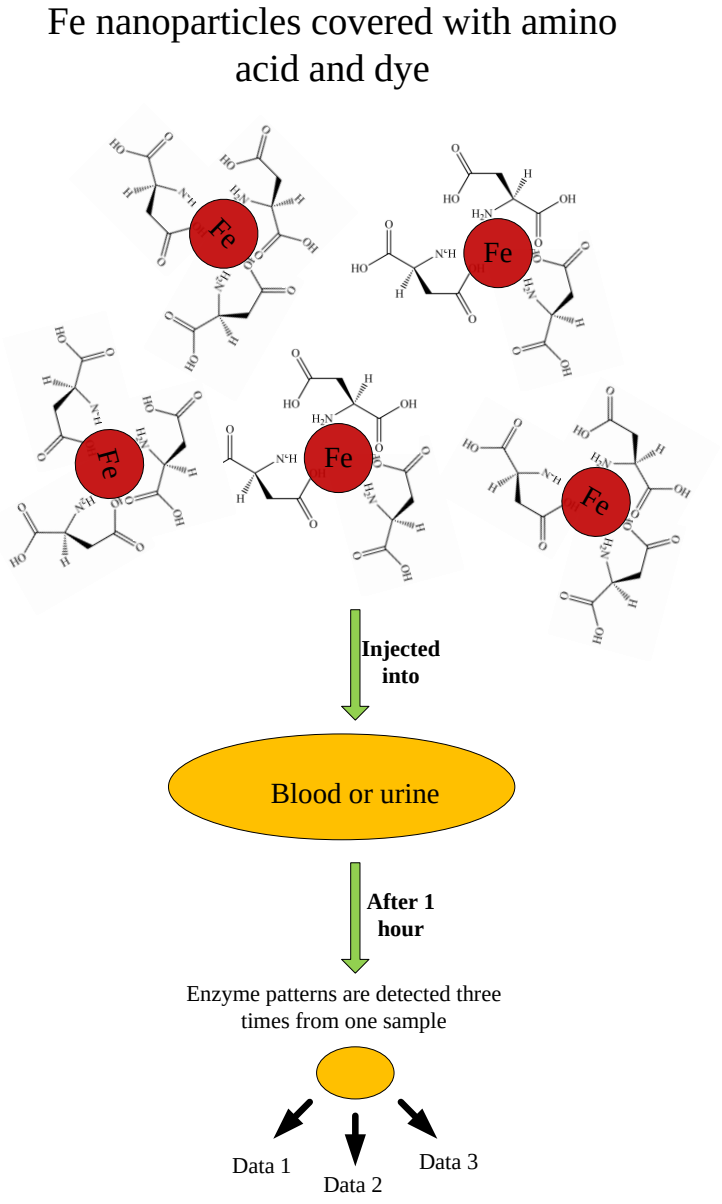
Data for this project were obtained from an experiment conducted by Dr. Bossmann and his coworkers in the Chemistry Department at Kansas State University. This experiment is aimed at testing the ability of a newly-developed nanoparticle in detecting cancer at an early stage by measuring enzyme activity in individuals. A small amount of blood or urine is obtained and iron nanoparticles coated with amino acids and a dye are added. The amino acids and dye can react with enzymes in the patients. Different types of cancers or different stages of cancer could lead to different enzyme patterns. The enzyme pattern may then be used to identify cancer by doctors. This test can also be used to distinguish between cancers that commonly occur in the human body, and monitor cancer in the treatment process. The detection process is roughly 60 minutes, and is supposed to be reduced to 5 minutes.

The raw data for this project include patients' cancer stage, age, and a number for enzyme activity. The aim of this project is to investigate methods for predicting stage of cancer for any patient based on his or her enzyme activity in the test.

1.2 Measurement procedure

The measurement procedure is represented in Figure 1-1. For each test, there are two groups of individuals, a patient group and a control group. Iron nanoparticles covered with amino acid and a dye were added into urine or blood of all individuals. After 1 hour, enzyme patterns were measured three times from one sample.

Figure 1-1 Procedure to test enzyme activity in individuals



1.3 Overview of report

In this report, three datasets related to the above diagnostic test are analyzed using two statistical methods, logistic regression and classification and regression tree (CART). The aim is to find a good method to predict the stage of cancer for a patient based on his or her enzyme activity. For each analysis, a measure of sensitivity and specificity for the test is estimated. These in turn are used to produce a receiver operation characteristic (ROC)

curve to determine an optimal cutoff to predict cancer or no cancer. Chapter 2 is a literature review on logistic regression and CART. Chapter 3 and Chapter 4 discuss how to use these methods to analyze the datasets collected for patients with breast cancer using a CathB nanoparticle and a MP nanoparticle, respectively, to induce enzyme activity. Then in Chapter 5, the same method is used to analyze the dataset collected for patients with lung cancer using a MP nanoparticle to induce enzyme activity.

Chapter 2 - Literature Review

2.1 Common measures for assessing diagnostic test

Diagnostic tests are often used to diagnose or detect disease, for example, using a new instrument in identifying the presence of cancer. Sensitivity and specificity are the most commonly used statistical measures that evaluate the performance of a diagnostic test. [1] [2] Sensitivity is the proportion of true positives that are correctly identified by the test. Specificity is the proportion of true negatives that are correctly identified by the test. [1] True positive means the individual tested has a disease or a symptom. A perfect predictor for the binary classification test would lead to 100 % sensitivity and 100 % specificity. But in reality, sensitivity and specificity will always be less than 100 %. When a new test method is proposed, these two measures are always reported.

$$\text{Specificity} = \frac{number\ of\ patients\ with\ abnormal\ pathology\ that\ are\ correctly\ identified}{number\ of\ patients\ with\ abnormal\ pathology}$$

$$\text{Sensitivity} = \frac{number\ of\ patients\ without\ abnormal\ pathology\ that\ are\ correctly\ identified}{number\ of\ patients\ without\ abnormal\ pathology}$$

Besides sensitivity and specificity, it is also necessary to know how good the test is at predicting an abnormality. The positive predictive value (PPV) and negative predictive value (NPV) are used for this purpose. [2] PPV is the proportion of patients with positive test results who are correctly diagnosed. NPV is the proportion of patients with negative test results who are correctly diagnosed. [2] A good test should have a high PPV and a high NPV. But PPV and NPV are functions of sensitivity, specificity and prevalence; even if the sensitivity and specificity are high, PPV will be much lower than 100 % when the prevalence of the disease is low. Formulas for PPV, NPV and prevalence are as follows.

$$\text{PPV} = \frac{sensitivity * prevalence}{sensitivity * prevalence + (1 - specificity) * (1 - prevalence)}$$

$$\text{NPV} = \frac{specificity * prevalence}{(1 - sensitivity) * prevalence + specificity * (1 - prevalence)}$$

$$\text{Prevalence of abnormality} = \frac{number\ of\ abnormal\ individuals}{number\ of\ individuals\ in\ the\ test}$$

A likelihood ratio is another number which is used to assess the performance of a test. It is calculated by sensitivity/ (1-specificity). [3] A likelihood ratio is used to compare the probability of getting positive result if the patient truly had the condition of interest with the corresponding probability if the patient was healthy. A high likelihood ratio is desired and can show that the test is useful. [3]

$$\text{Likelihood ratio} = \frac{sensitivity}{(1 - specificity)}$$

2.2 Introduction to statistical analysis methods used in this project

2.2.1 Introduction to the logistic regression model

Regression models are an important part of data analysis which involves describing the relationship between a response and explanatory response variables. When the response has one or several possible values, logistic regression is a common method. The purpose of logistic regression is to find the best fitting and biologically reasonable model for the relationship between a response variable and explanatory variables. For a logistic regression, the random component is the condition of the response, either success or failure. [5] [6]

Maximum likelihood is the method for estimation of parameters in a logistic regression model. This method gives estimates for the unknown parameters which maximize the probability of obtaining the observed set of data. [7] When the response is binary, either failure or success, a logistic model is fit to estimate the probability of success given a fixed value of an explanatory variable or fixed values of several explanatory variables, denoted as $P(Y = 1|x)$. [7] The format of logistic regression equation is

$$\text{logit } \pi(x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

where $\pi(x)$ is the probability of success, α is the intercept of the model, β_i is the coefficient for the ^{i}th predictor. Rearrangement of this fitted model gives the estimated probability of success

$$\widehat{\pi(x)} = \frac{e^{\hat{\alpha}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\cdots}}{1 + e^{\hat{\alpha}+\widehat{\beta_1}x_1+\widehat{\beta_2}x_2+\cdots}}$$

For each individual, the 95% confidence interval for a probability is computed using the following two steps, assuming one explanatory variable.

Step 1 Make a confidence interval for logit(π(x)): (l, u)

$$\mathrm{SE}(\hat{\alpha}+\hat{\beta}\mathrm{x}) = \sqrt{var\ (\hat{\alpha}) + x^2 var(\hat{\beta}) + 2cov(\hat{\alpha},\hat{\beta})}$$

$$(\hat{\alpha}+\hat{\beta}\mathrm{x}) \pm Z_{\alpha/2}\mathrm{SE}(\hat{\alpha}+\hat{\beta}\mathrm{x})$$

Step 2 Make a confidence interval

$$(\frac{e^l}{1+e^l}, \frac{e^u}{1+e^u}) \text{ for } \pi(\mathrm{x})$$

When the response variables have more than two categories, several logit models can be used. If there are J possible values for the response variable, a set of J-1 equations should be fit.

π_1=probability that y is in category 1

π_2=probability that y is in category 2

π_J=probability that y is in the last category and is used as the baseline

The model is $\log\frac{\pi_j}{\pi_J} = \alpha_j + \beta_j x$, if only one x variable is in the model.

$$\widehat{\pi_j} = \frac{\exp(\widehat{\alpha_j} + \widehat{\beta_j}x)}{\sum_{h=1}^{J}\exp(\widehat{\alpha_h} + \widehat{\beta_h}x)}$$

where the parameters for the last baseline category are 0, i.e., $\alpha_J = \beta_J = 0$.

$$\pi_J = 1 - \pi_1 - \pi_2 \cdots - \pi_{J-1}$$

The goodness-of-fit assessment in a logistic regression model is by a statistic called deviance, denoted by D. [5] Deviance is used to test if the current model is appropriate. The expression for deviance is

$$\mathrm{D} = -2\ln[\frac{likelihood\ of\ the\ current\ model}{likelihood\ of\ the\ saturated\ model}]$$

In order to test the significance of the coefficients, one needs to compare the value of D with and without the variable in question. [5]

$$\mathrm{G} = \mathrm{D}(\text{for the model without the variable}) - \mathrm{D}(\text{for the model with the variable})$$

$$= -2\ln\left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right]$$

The statistic G asymptotically follows a chi-square distribution with 1 degree of freedom under the hypothesis that the coefficient for the tested variable is 0. As a result, a chi-square test can be used to approximately test the significance of coefficients for different explanatory variables.

2.2.2 ROC curves

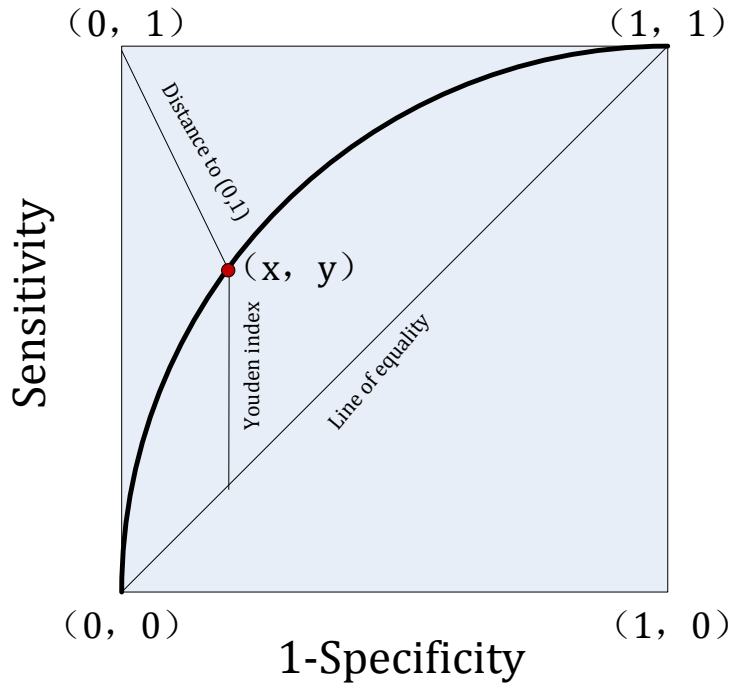
For quantitative tests, there must be a way to assess the accuracy of the test and there should be a decision threshold of probability in order to provide a prediction of success/false. The sensitivity and specificity of a test will be changed with the decision threshold accordingly.

A receiver operating characteristic (ROC) curve can be used to assess the accuracy of the diagnostic test and to give a cut-off probability for the prediction of the response. [8] A ROC curve summarizes predictive power for all possible thresholds and is a plot of sensitivity as a function of (1-specificity) for the possible threshold or cut-off probability. It measures the ability of the test to correctly classify those with and without the disease. The curve commonly has a concave shape and connects the points (0,0) and (1,1). [9] The area under the ROC curve is the same as the concordance index which is the value of a measure of predictive power. [10] The higher the area, the greater the predictive power of the test. If the concordance index is 0.5, the predictions were no better than random guessing. The accuracy of a diagnostic test is classified following a conventional guide: if the area is from 0.9 to 1, the accuracy of the test is excellent; if the area is from 0.8 to 0.9, the accuracy of the test is good; if the area is from 0.7 to 0.8, the accuracy of the test is fair; if the area is 0.6 to 0.7, the accuracy of the test is poor; if the area is from 0.5 to 0.6, the accuracy of the test fails. [10] But an ROC curve can only be used for a test with binary responses.

There are three criteria to find an optimal threshold of probability from the ROC curve. [9] The first two methods give equal weight to sensitivity and specificity and consider no ethical or cost constraints. The third criterion considers financial cost for correct and false diagnoses, cost of discomfort to a person caused by treatment, and cost of further investigation. The third method is seldom used because of its complexity. In this literature

review, only the first two methods are discussed. Shown in Figure 2-1 is a ROC curve. The first method computes the distance from each point on the ROC curve to the point (0, 1). The point with the shortest distance is the optimal threshold. The second method computes the vertical distance from each point on the ROC curve to the line of equality and the point with the maximum vertical distance is the optimal threshold. The vertical distance is called the Youden index.

Figure 2-1 Methods to find the best cut-off from the ROC curve



2.2.3 Introduction to classification and regression trees

The method of classification and regression trees (CART) is used to select variables and their interactions which are important in determining an outcome. [11; 12] CART can also be used to classify statistical data. If the dependent variable is continuous, CART produces a regression tree. If the dependent variable is categorical, CART produces a classification tree. The purpose of CART is to find a set of classifiers which are responsible for a given phenomenon. Using these classifiers, it is possible that the outcome of a new observation can be predicted by these classifiers. The steps in the tree-building process involve (1) building a large tree with many nodes, (2) combining some of the branches to generate

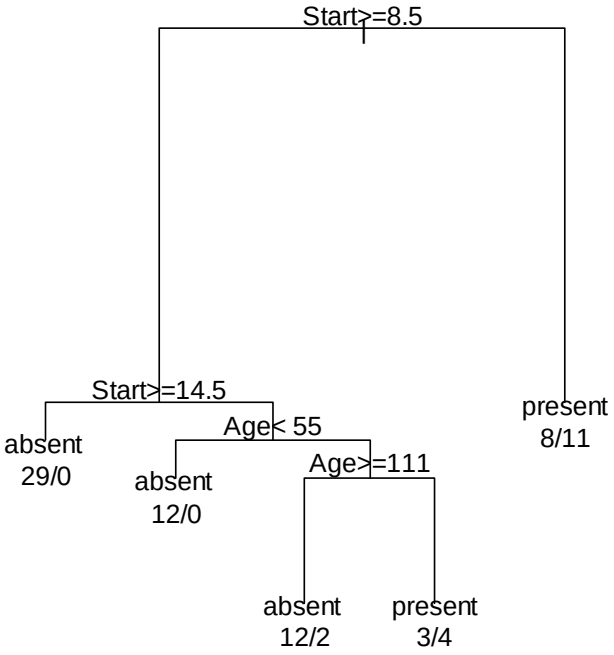
different subtrees, and (3) choosing the best tree among all the subtrees. The split of a CART node begins with a simple question requiring a yes/no answer, such as is the value of a variable, Xo, less than or greater than a particular value. Based on the answer, the sample is split into left and right subsamples. Then the subsamples are split further based on other questions until the subsamples are homogeneous or contain two few observations. [11; 12; 13] The variable that is chosen to divide the dataset into two groups, and its value, is determined by the longest reduction in deviance from the parent mode to the two children (subsample) nodes. With a continuous response variable, the deviance of a node is the sums of squared derivations about the mean of that node. In the case of binary data, the deviance is that used for binomial data.

In this literature review, a dataset called Kyphosis is used to illustrate how the CART model is used to build a tree. This dataset involves observations on 81 children undergoing corrective surgery of the spine. In the dataset, there are three risk factors for kyphosis after the surgery, including age in months (called Age), the starting vertebral level of the surgery (called Start), and the number of vertebrae involved (called Number). A classification and regression tree is used to predict if a child has kyphosis after the surgery based on the levels of the three risk factors. The tree is shown in Figure 2-2. The tree is built starting with a question if the age of the child is greater or equal to 8.5 months. If the answer to the question is yes, this child is partitioned to the left subsamples. If the answer is no, the child is portioned to the right of the subsamples and predicted to have kyphosis. Splitting the original dataset based on whether the child was greater or less than 8.5 months obtained the greatest reduction in deviance versus all other possible splits. The split of the left subsamples is continued with the second question if the starting vertebral level of the surgery is greater or equal to 14.5. If the answer is yes, the child is predicted to not have kyphosis. If the answer is no, this child is partitioned to the right sub-subsamples. The sub-subsamples are split further based on other questions with the goal of each split to improve the accuracy of predictions, i.e., to obtain homogenous nodes. There are two numbers under each node in Figure 2-2. The first number represents out of the total individuals in the subsample, how many of them do not have kyphosis. The second number represents how many of them have kyphosis. For example, in the right most node, which is labeled

present, 8 children did not develop kyphosis after the surgery, whereas, 11 children developed kyphosis after the surgery.

The strengths of CART include (1) no assumption of the distributions for the variables is needed, (2) the explanatory variables can be either continuous, or categorical, or a combination of both, (3) outliers do not affect the result as much as other modeling techniques, (4) transformation of explanatory variables does not have an effect on the tree. CART can analyze the data and reveal the importance of each explanatory variable. CART also has some weaknesses. For example, the split of data into subgroups is only based on a single explanatory variable, and there is no probability level or confidence interval for predictions obtained from a CART tree. [11; 12]

Figure 2-2 A classification and regression tree built on dataset of kyphosis [13]



Chapter 3 - Assessing Diagnostic Test for Breast Cancer (CathB)

3.1 Overview of the data

There are two groups in the dataset. The first group includes 20 patients with breast cancer at different stages (0, I, II, III, IV), and the second group includes 12 individuals who do not have cancer. The second group is called the control group. Each individual was asked to provide personal information, including age and stage of cancer. The nanoparticles added to the blood samples are coated with Cathepsin B, which is a type of protein to interact with the enzyme in the blood samples. The level of CathB (enzyme activity) was tested three times. The average level of CathB is computed and used as the random variable to predict level of breast cancer. Relationships between any two variables are displayed in Figure 3-1, Figure 3-2 and Figure 3-3. From these boxplots, it can be seen that patients with cancer stages of III and IV have higher average CathB than the persons without breast cancer or the patients with breast cancer of stage 0, I, and II. Age of patient does not show a significant relationship with average CathB. But age of patient seems to have a relationship with stages of breast cancer; patients with breast cancer of stage I, II, III, and IV are older than the persons without breast cancer or with breast cancer of stage 0.

Figure 3-1 Relationship between average CathB and staging of cancer

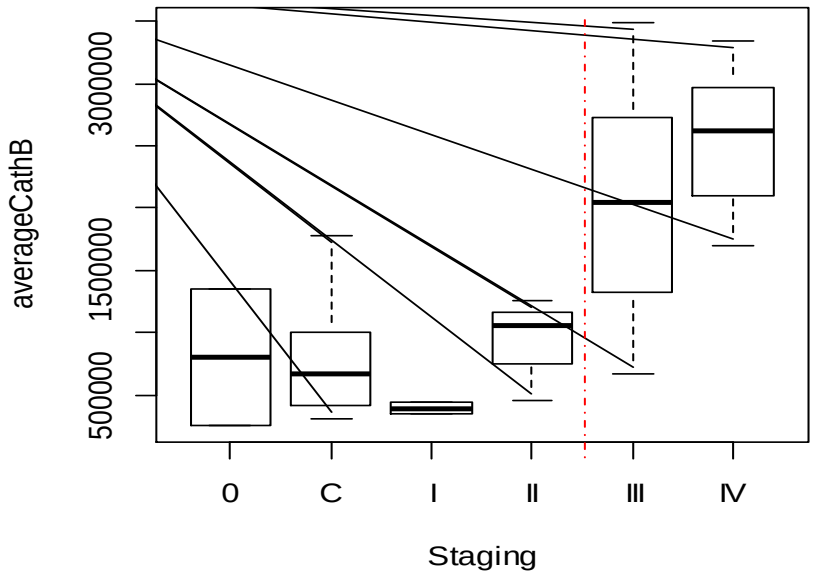


Figure 3-2 Relationship between average CathB and age

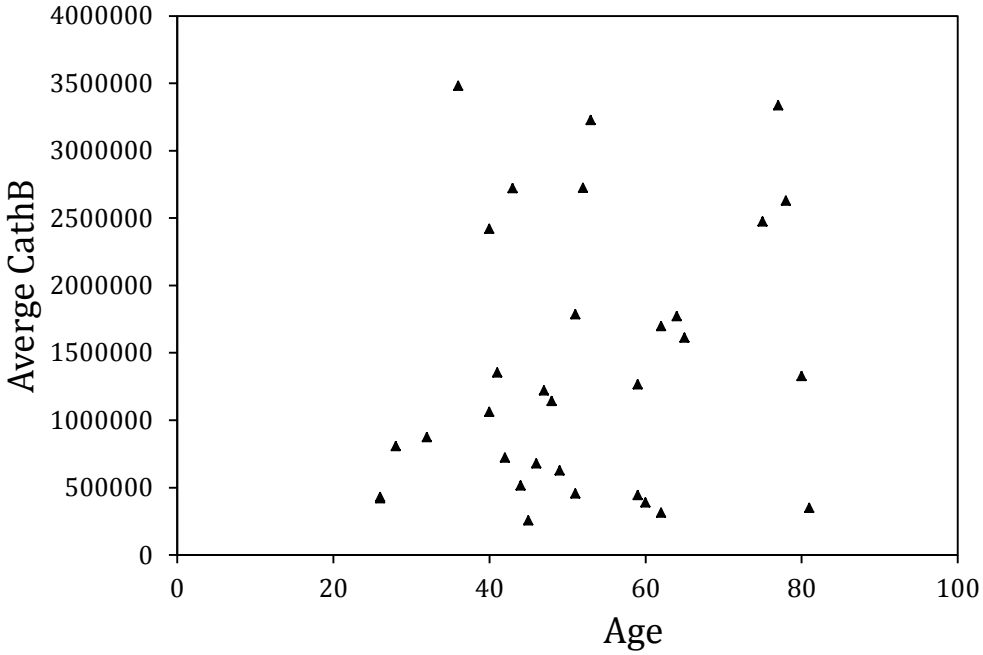
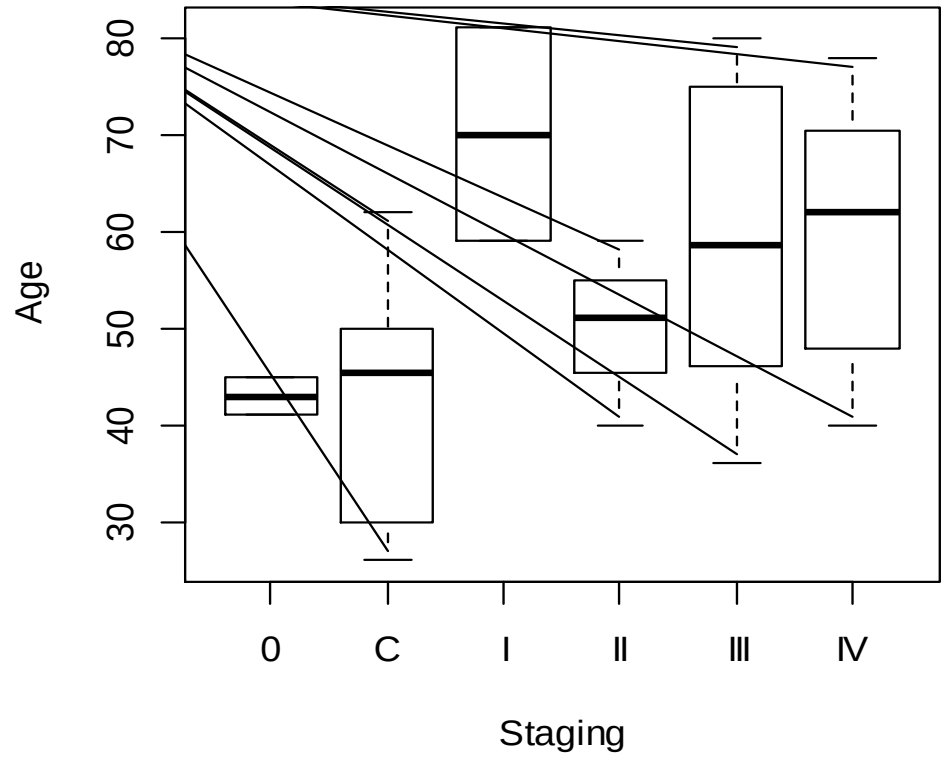


Figure 3-3 Relationship between age and staging of cancer



3.2 Comparison of three measurements

The enzyme pattern of each person was tested three times on different days. Figure 3-4 shows the results of the three measurements. It can be seen that the experimental error is very small. In addition, a logistic regression model is used to predict the probability of having severe breast cancer for each person by using the individual enzyme pattern instead of the average enzyme pattern of the three measurements. Severe breast cancer includes breast cancer of stages III and IV. Moderate breast cancer includes breast cancer of stages 0, I, II and no breast cancer. The reason why the patients are grouped into "moderate breast cancer" group and "severe breast cancer" group is because from the boxplot of breast cancer staging vs enzyme pattern, it can be seen that patients having breast cancer of stage III and IV have higher enzyme activity than patients having breast cancer of stage 0, I, and II. This grouping also allows use of all observations in this initial analysis to investigate the consistency of the three measurements. Later, other groupings will be used when evaluating the sensitivity of the test for distinguishing stages of cancer. In addition, logistic regression requires the response to be binary. So the patients are separated into two groups. The probabilities of having severe breast cancer for each patient are listed in Table 3-1, showing that there is no significant difference between the three probabilities. The three probabilities for each person are also shown in Figure 3-5. As a result, in this report, the models are built based on the average value of three measurements.

Figure 3-4 Comparison of the three enzyme patterns of different individuals

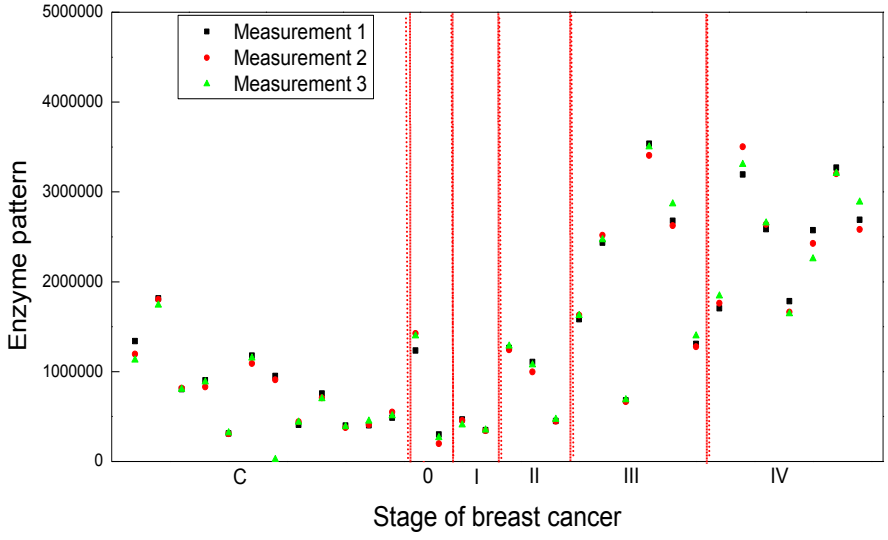
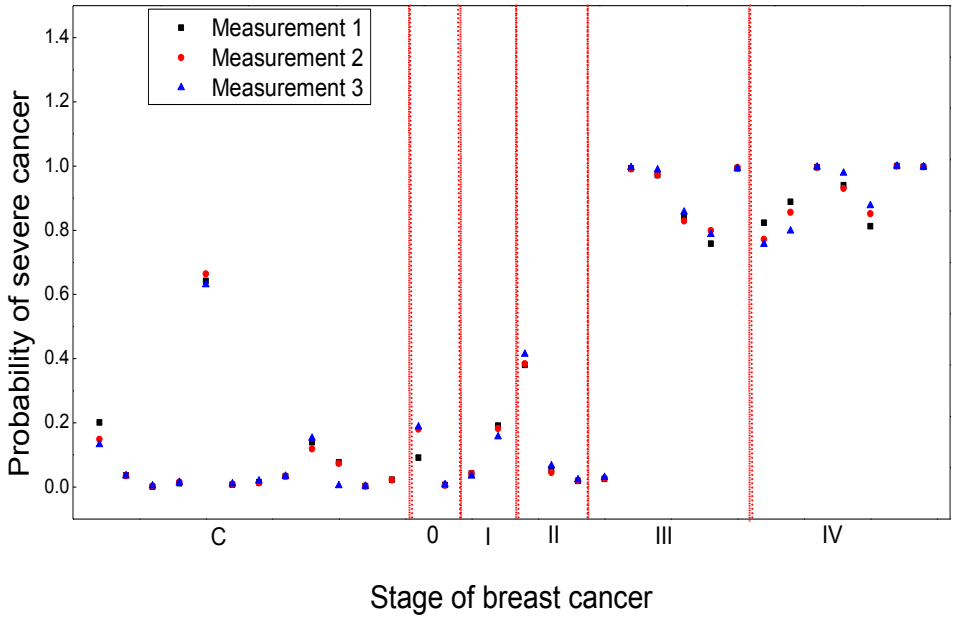


Table 3-1 Probabilities of severe cancer for each patient predicted by logistic regression, calculated separately for three measurements

Obs	Patient	Age	Staging	level	P1	P2	P3
1	B1	41	0	moderate	0.091	0.180	0.187
2	B10	59	II	moderate	0.380	0.384	0.414
3	B11	46	III	severe	0.025	0.026	0.030
4	B12	62	IV	severe	0.823	0.772	0.756
5	B13	40	IV	severe	0.888	0.856	0.798
6	B14	40	II	moderate	0.056	0.045	0.066
7	B15	36	III	severe	0.993	0.991	0.995
8	B16	53	IV	severe	0.996	0.996	0.997
9	B17	52	III	severe	0.972	0.972	0.988
10	B18	43	IV	severe	0.940	0.930	0.978
11	B19	51	II	moderate	0.019	0.020	0.022
12	B2	64	IV	severe	0.812	0.851	0.877
13	B20	80	III	severe	0.840	0.829	0.857
14	B3	45	0	moderate	0.006	0.005	0.007
15	B4	59	I	moderate	0.041	0.041	0.034
16	B5	65	III	severe	0.758	0.799	0.787
17	B6	77	IV	severe	1.000	1.000	1.000
18	B6a	81	I	moderate	0.191	0.182	0.156
19	B7	78	IV	severe	0.997	0.998	0.997
20	B8	75	III	severe	0.993	0.995	0.993
21	C1	47	C	moderate	0.201	0.148	0.132
22	C10	60	C	moderate	0.036	0.035	0.035
23	C11	26	C	moderate	0.001	0.002	0.003
24	C12	44	C	moderate	0.011	0.015	0.014
25	C2	51	C	moderate	0.642	0.664	0.630
26	C3	28	C	moderate	0.007	0.008	0.010
27	C4	32	C	moderate	0.014	0.012	0.019
28	C5	62	C	moderate	0.033	0.033	0.033
29	C6	48	C	moderate	0.139	0.118	0.152
30	C7	49	C	moderate	0.077	0.073	0.004
31	C8	26	C	moderate	0.002	0.002	0.002
32	C9	42	C	moderate	0.022	0.021	0.023

Figure 3-5 Scatterplot of severe cancer probabilities for each patient predicted by logistic regression using individual enzyme pattern as the predictor



3.3Analysis based on logistic regression with binary response (stage 0 and I deleted)

3.3.1 Use of average CathB as the predictor

A logistic regression model is used to predict the probability of having breast cancer for each person. Patients with breast cancer of staging II, III, and IV are grouped together as having breast cancer, whereas the persons without breast cancer are grouped together as having no breast cancer. Patients with breast cancer of stage 0 or I are deleted because they have very low enzyme activity although they have breast cancer. By deleting these stages, the model will have less error and has a better prediction on the existence of breast cancer at later stages II, III, and IV. There are only 2 persons having breast cancer of stage 0 and 2 persons having breast cancer of stage I. Only 4 persons are deleted. When average CathB is used as the predictor, the estimates of parameters are shown in Table 3-2. Probability of having breast cancer for each patient is shown in Table 3-3.

Table 3-2 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8809	1.1731	6.0314	0.0141
CathB	1	2.532E-6	9.857E-7	6.5954	0.0102

Table 3-3 Probability of having breast cancer for each patient

Obs	Patient	Age	Staging	CathB	Cancer	Prob
1	C1	47	C	1220798	No	0.552
2	C2	51	C	1785704	No	0.838
3	C3	28	C	806983.7	No	0.302
4	C4	32	C	873253.3	No	0.339
5	C5	62	C	313112.3	No	0.110
6	C6	48	C	1140311	No	0.502
7	C7	49	C	627024.3	No	0.215
8	C8	26	C	428675.7	No	0.142
9	C9	42	C	720771.7	No	0.258
10	C10	60	C	387576	No	0.130
11	C11	26	C	420150.7	No	0.140
12	C12	44	C	515214.7	No	0.171
13	B10	59	II	1264716	Yes	0.580
14	B14	40	II	1060055	Yes	0.451
15	B19	51	II	456155	Yes	0.151
16	B5	65	III	1611268	Yes	0.768
17	B8	75	III	2473030	Yes	0.967
18	B11	46	III	677698.3	Yes	0.238
19	B15	36	III	3481133	Yes	0.997
20	B17	52	III	2723205	Yes	0.982
21	B20	80	III	1327514	Yes	0.618
22	B2	64	IV	1770087	Yes	0.832
23	B6	77	IV	3335243	Yes	0.996
24	B7	78	IV	2626708	Yes	0.977
25	B12	62	IV	1697287	Yes	0.805
26	B13	40	IV	2419770	Yes	0.963
27	B16	53	IV	3226621	Yes	0.995
28	B18	43	IV	2720022	Yes	0.982

The optimal cut-off probability to predict if a person has breast cancer is found by a ROC curve. The ROC curve is obtained by plotting sensitivity and 1-specificity for different cut-off values. Table 3-4 lists sensitivity and 1-specificity calculated under different cut-off values and Figure 3-6 is the ROC curve. The optimal point is the one which has the smallest distance to the point (0, 1) and at the same time has the largest vertical distance to the line of equality. Based on these criteria, the probability of 0.580 is the optimal cut-off

probability to predict if the person has breast cancer. If the predicted probability of a person is above 0.580, this person is predicted to have breast cancer. If the predicted probability of a person is below 0.580, this person is predicted to have no breast cancer. The area under the ROC curve is 0.8854, indicating this test method is good, but not excellent.

Figure 3-6 ROC curve of the model

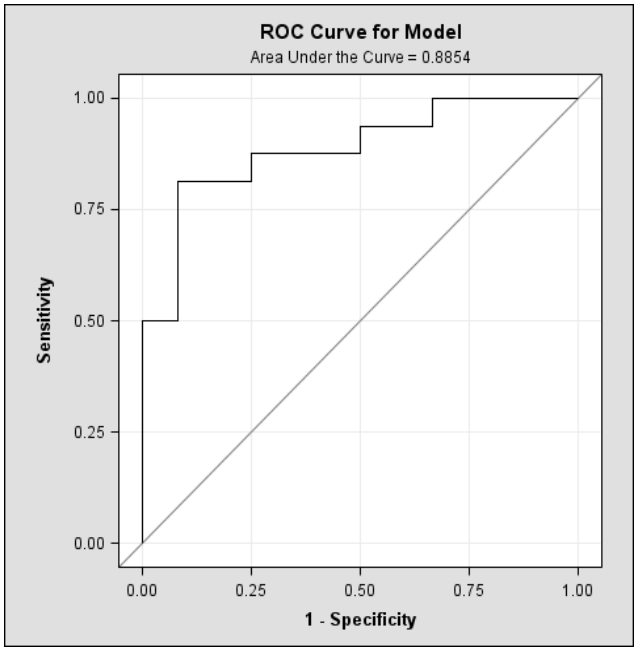


Table 3-4 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	0.997	0.06	0.00	0.94	0.06
2	0.996	0.13	0.00	0.88	0.13
3	0.995	0.19	0.00	0.81	0.19
4	0.982	0.25	0.00	0.75	0.25
5	0.982	0.31	0.00	0.69	0.31
6	0.977	0.38	0.00	0.63	0.38
7	0.967	0.44	0.00	0.56	0.44
8	0.962	0.50	0.00	0.50	0.50
9	0.837	0.50	0.08	0.51	0.42
10	0.832	0.56	0.08	0.45	0.48
11	0.805	0.63	0.08	0.38	0.54
12	0.768	0.69	0.08	0.32	0.60
13	0.618	0.75	0.08	0.26	0.67
14	0.580	0.81	0.08	0.21	0.73

15	0.552	0.81	0.17	0.25	0.65
16	0.501	0.81	0.25	0.31	0.56
17	0.451	0.88	0.25	0.28	0.63
18	0.338	0.88	0.33	0.36	0.54
19	0.302	0.88	0.42	0.44	0.46
20	0.258	0.88	0.50	0.52	0.38
21	0.238	0.94	0.50	0.50	0.44
22	0.215	0.94	0.58	0.59	0.35
23	0.171	0.94	0.67	0.67	0.27
24	0.151	1.00	0.67	0.67	0.33
25	0.142	1.00	0.75	0.75	0.25
26	0.140	1.00	0.83	0.83	0.17
27	0.130	1.00	0.92	0.92	0.08
28	0.110	1.00	1.00	1.00	0.00

When the cut-off probability is set to be 0.580, the relationship between the predicted existence of breast cancer and the actual diagnosis is shown in Figure 3-7. One person without breast cancer is predicted to have breast cancer, whereas, three persons with breast cancer are predicted to have no breast cancer. The sensitivity and specificity for this test is 0.81 and 0.92, respectively. Table 3-5 summarizes how many persons are predicted to have breast cancer and how many persons are predicted to have no breast cancer for patients in different stages of breast cancer.

Figure 3-7 Prediction of existence of cancer based on the optimal cut-off probability

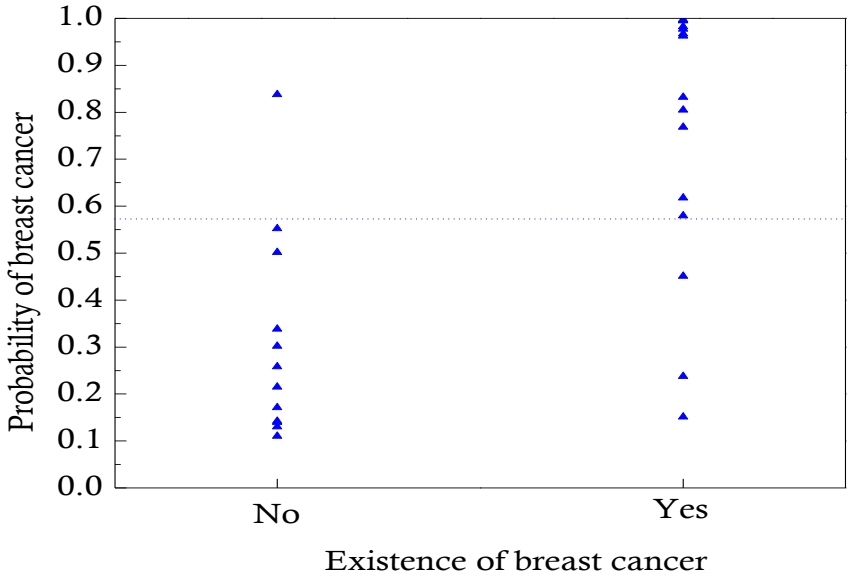


Table 3-5 Prediction of existence of breast cancer for patients in different stages of breast cancer

	Predicted	
	No	Yes
C	11	1
II	2	1
III	1	5
IV	0	7

3.3.2 Use of average CathB and age as the predictor

In this section, average CathB and age are used as the predictors to predict if the person has breast cancer. Estimates of parameters are shown in Table 3-6. The probability of having breast cancer for each patient is shown in Table 3-7.

Table 3-6 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.2938	2.8811	4.7720	0.0289
CathB	1	2.073E-6	8.738E-7	5.6290	0.0177
Age	1	0.0785	0.0536	2.1440	0.1431

Table 3-7 Probability of having breast cancer for each patient

Obs	Patient	Age	Staging	CathB	Cancer	Prob
1	B2	64	IV	1770087	Yes	0.917
2	B5	65	III	1611268	Yes	0.896
3	B6	77	IV	3335243	Yes	0.999
4	B7	78	IV	2626708	Yes	0.995
5	B8	75	III	2473030	Yes	0.991
6	B10	59	II	1264716	Yes	0.723
7	B11	46	III	677698.3	Yes	0.218
8	B12	62	IV	1697287	Yes	0.890
9	B13	40	IV	2419770	Yes	0.866
10	B14	40	II	1060055	Yes	0.278
11	B15	36	III	3481133	Yes	0.977
12	B16	53	IV	3226621	Yes	0.990
13	B17	52	III	2723205	Yes	0.969
14	B18	43	IV	2720022	Yes	0.938
15	B19	51	II	456155	Yes	0.207
16	B20	80	III	1327514	Yes	0.939
17	C1	47	C	1220798	No	0.482

18	C2	51	C	1785704	No	0.804
19	C3	28	C	806983.7	No	0.081
20	C4	32	C	873253.3	No	0.122
21	C5	62	C	313112.3	No	0.315
22	C6	48	C	1140311	No	0.460
23	C7	49	C	627024.3	No	0.241
24	C8	26	C	428675.7	No	0.033
25	C9	42	C	720771.7	No	0.182
26	C10	60	C	387576	No	0.314
27	C11	26	C	420150.7	No	0.033
28	C12	44	C	515214.7	No	0.145

Table 3-8 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 3-8 is the ROC curve. From Table 3-8, it is found that 0.723 is the optimal cut-off probability to predict if a person has breast cancer or not. If the predicted probability of a person is above 0.723, this person is predicted to have breast cancer. If the predicted probability of a person is below 0.723, this person is predicted to have no breast cancer. The area under the ROC curve is 0.9063, indicating this test method is excellent.

Figure 3-8 ROC curve of the model

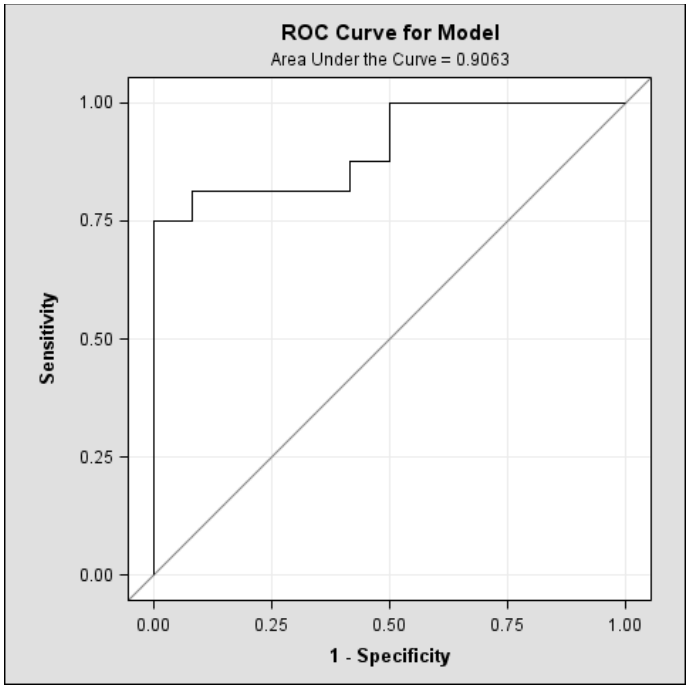


Table 3-8 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

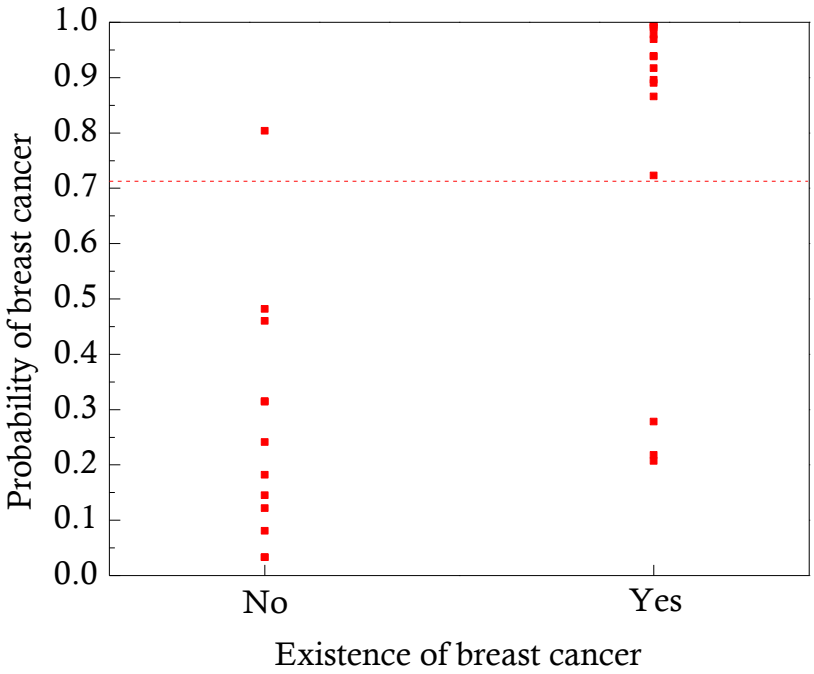
Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	0.999	0.06	0.00	0.94	0.06
2	0.995	0.13	0.00	0.88	0.13
3	0.991	0.19	0.00	0.81	0.19
4	0.990	0.25	0.00	0.75	0.25
5	0.977	0.31	0.00	0.69	0.31
6	0.969	0.38	0.00	0.63	0.38
7	0.939	0.44	0.00	0.56	0.44
8	0.938	0.50	0.00	0.50	0.50
9	0.917	0.56	0.00	0.44	0.56
10	0.895	0.63	0.00	0.38	0.63
11	0.890	0.69	0.00	0.31	0.69
12	0.865	0.75	0.00	0.25	0.75
13	0.804	0.75	0.08	0.26	0.67
14	0.723	0.81	0.08	0.21	0.73
15	0.481	0.81	0.17	0.25	0.65
16	0.459	0.81	0.25	0.31	0.56
17	0.314	0.81	0.33	0.38	0.48
18	0.314	0.81	0.42	0.46	0.40
19	0.277	0.88	0.42	0.44	0.46
20	0.241	0.88	0.50	0.52	0.38
21	0.218	0.94	0.50	0.50	0.44
22	0.207	1.00	0.50	0.50	0.50
23	0.182	1.00	0.58	0.58	0.42
24	0.145	1.00	0.67	0.67	0.33
25	0.122	1.00	0.75	0.75	0.25
26	0.081	1.00	0.83	0.83	0.17
27	0.033	1.00	0.92	0.92	0.08
28	0.033	1.00	1.00	1.00	0.00

When the cut-off probability is set to be 0.723, the relationship between the predicted existence of breast cancer and the actual diagnosis is shown in Figure 3-9. One person without breast cancer is predicted to have breast cancer, whereas, three persons with breast cancer are predicted to have no cancer. The sensitivity and specificity for this test is 0.81 and 0.92, respectively. Table 3-9 summarizes how many persons are predicted to have breast cancer and how many persons are predicted to have no breast cancer for patients at different stages of breast cancer.

Table 3-9 Prediction of existence of breast cancer for patients at different stages

	Predicted	
	No	Yes
C	11	1
II	2	1
III	1	5
IV	0	7

Figure 3-9 Prediction of existence of cancer based on the optimal cut-off probability



3.3.3 Comparison of models in section 3.3.1 and 3.3.2

Table 3-10 compares the probabilities calculated by the models in section 3.3.1 (average CathB is used as the predictor) and section 3.3.2 (average CathB and age are used as the predictors). The probabilities do not have a big difference for most of the persons. The maximum difference is 0.321, showing up on patient B20 with a cancer stage III. If the age of a patient is close to the average age of the persons involved in the test, the difference between the two probabilities is smaller. If the age of a patient is very different from the average age, the difference between the two probabilities is big.

Table 3-10 Comparison of probabilities of having breast cancer for each patient between the two models in section 3.3.1 and 3.3.2

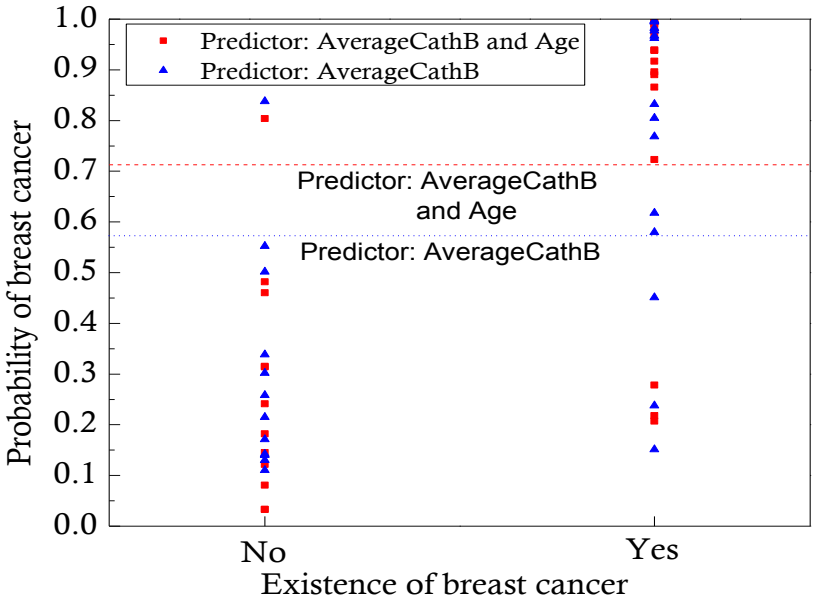
Obs	Patient	Age	Staging	CathB	Cancer	Prob1	Prob2
13	B10	59	II	1264716	Yes	0.580	0.723
18	B11	46	III	677698.3	Yes	0.238	0.218
25	B12	62	IV	1697287	Yes	0.805	0.890
26	B13	40	IV	2419770	Yes	0.963	0.866
14	B14	40	II	1060055	Yes	0.451	0.278
19	B15	36	III	3481133	Yes	0.997	0.977
27	B16	53	IV	3226621	Yes	0.995	0.990
20	B17	52	III	2723205	Yes	0.982	0.969
28	B18	43	IV	2720022	Yes	0.982	0.938
15	B19	51	II	456155	Yes	0.151	0.207
22	B2	64	IV	1770087	Yes	0.832	0.917
21	B20	80	III	1327514	Yes	0.618	0.939
16	B5	65	III	1611268	Yes	0.768	0.896
23	B6	77	IV	3335243	Yes	0.996	0.999
24	B7	78	IV	2626708	Yes	0.977	0.995
17	B8	75	III	2473030	Yes	0.967	0.991
1	C1	47	C	1220798	No	0.552	0.482
10	C10	60	C	387576	No	0.130	0.314
11	C11	26	C	420150.7	No	0.140	0.033
12	C12	44	C	515214.7	No	0.171	0.145
2	C2	51	C	1785704	No	0.838	0.804
3	C3	28	C	806983.7	No	0.302	0.081
4	C4	32	C	873253.3	No	0.339	0.122
5	C5	62	C	313112.3	No	0.110	0.315
6	C6	48	C	1140311	No	0.502	0.460
7	C7	49	C	627024.3	No	0.215	0.241
8	C8	26	C	428675.7	No	0.142	0.033
9	C9	42	C	720771.7	No	0.258	0.182

Table 3-11 summarizes the differences of the two models in the diagnostic test of breast cancer. Including age as a predictor increases the area under the ROC curve, increases the cut-off probability, but does not change sensitivity and specificity of the test. The p-value for predictor average CathB is increased. The p-value for the predictor age is 0.1431, indicating that age has a marginal influence on the prediction. Comparison of predicted accuracy between the two models is shown in Figure 3-10.

Table 3-11 Comparison of models in section 3.3.1 and 3.3.2

Predictor	Area under the ROC curve	Threshold	Sensi	Speci	P-value	
					MMP	Age
Average CathB	0.8854	0.580	0.81	0.92	0.0102	N/A
Average CathB+Age	0.9063	0.723	0.81	0.92	0.0177	0.1431

Figure 3-10 Prediction of existence of breast cancer based on the optimal cut-off probability for the models used in section 3.3.1 and 3.3.2



3.4 Analysis based on logistic regression with binary response (stage II deleted)

A logistic regression is used to predict the probability of severe breast cancer based on average CathB and age. Patients at stage III and IV are group together and their breast cancer condition is severe. All the other individuals belong to the moderate group(control and stage 0 and I). The reason for division of patients into such two groups is explained in section 3.2. The individuals with breast cancer of stage II are deleted. The reason why they are deleted is because breast cancer of stage II is more advanced, but patients having breast cancer of stage II have much lower enzyme activity than the patients having breast cancer of stage III and IV. By deleting stage II, the model will have less error and has a better prediction on the existence of severe breast cancer. There are only 3 persons having breast

cancer of stage II. Then the number of individuals in this analysis becomes 29. The different groupings done here in addition to what was done in earlier subsections should improve insight into the performance capabilities of the test in distinguishing stages of cancer. Confidence interval for the probabilities of having severe breast cancer are calculated in section 3.41 and used as an example to show how to calculate confidence intervals and how to use confidence intervals. Due to complexity, confidence intervals are not calculated in other sections.

3.4.1 Use of average CathB as the predictor

When average CathB is used as the predictor, the estimates of parameters are shown in Table 3-12. The predicted probability of severe breast cancer and 95% confidence interval of the probability are listed in Table 3-13.

Table 3-12 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8059	1.7494	7.5468	0.0060
CathB	1	3.381E-6	1.259E-6	7.2130	0.0072

Table 3-13 Probability of having severe breast cancer for each person

Obs	Patient	Staging	CathB	Level	Prob	L-Prob	U-Prob
1	C1	C	1220798	moderate	0.337	0.124	0.645
2	C2	C	1785704	moderate	0.774	0.370	0.952
3	C3	C	806983.7	moderate	0.111	0.022	0.416
4	C4	C	873253.3	moderate	0.135	0.030	0.445
5	C5	C	313112.3	moderate	0.023	0.002	0.265
6	C6	C	1140311	moderate	0.279	0.094	0.591
7	C7	C	627024.3	moderate	0.064	0.009	0.351
8	C8	C	428675.7	moderate	0.034	0.003	0.293
9	C9	C	720771.7	moderate	0.086	0.014	0.383
10	C10	C	387576	moderate	0.029	0.002	0.282
11	C11	C	420150.7	moderate	0.033	0.003	0.291
12	C12	C	515214.7	moderate	0.045	0.005	0.317
13	B1	0	1353355	moderate	0.443	0.181	0.740
14	B3	0	254411	moderate	0.019	0.001	0.251
15	B4	I	442882.7	moderate	0.035	0.003	0.297
16	B6a	I	347228.3	moderate	0.026	0.002	0.273
17	B5	III	1611268	severe	0.655	0.299	0.894
18	B8	III	2473030	severe	0.972	0.591	0.999

19	B11	III	677698.3	severe	0.075	0.011	0.368
20	B15	III	3481133	severe	0.999	0.802	1.000
21	B17	III	2723205	severe	0.988	0.654	1.000
22	B20	III	1327514	severe	0.421	0.170	0.722
23	B2	IV	1770087	severe	0.765	0.364	0.949
24	B6	IV	3335243	severe	0.998	0.779	1.000
25	B7	IV	2626708	severe	0.983	0.631	1.000
26	B12	IV	1697287	severe	0.718	0.335	0.928
27	B13	IV	2419770	severe	0.967	0.576	0.998
28	B16	IV	3226621	severe	0.998	0.760	1.000
29	B18	IV	2720022	severe	0.988	0.654	1.000

The optimal cut-off probability to predict if the person has severe breast cancer is found by a ROC curve. Table 3-14 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 3-11 is the ROC curve. The optimal point is the one which has the smallest distance to the point (0, 1) and at the same time has the largest vertical distance to the line of equality. Based on these criteria, the probability of 0.421 is the optimal cut-off probability to predict if the person has severe breast cancer. If the predicted probability of a person is above 0.421, this person is predicted to have severe breast cancer. If the predicted probability of a person is below 0.421, this person is predicted to have moderate breast cancer. The area under the ROC curve is 0.9423, indicating this test method is excellent.

Figure 3-11 ROC curve of the model

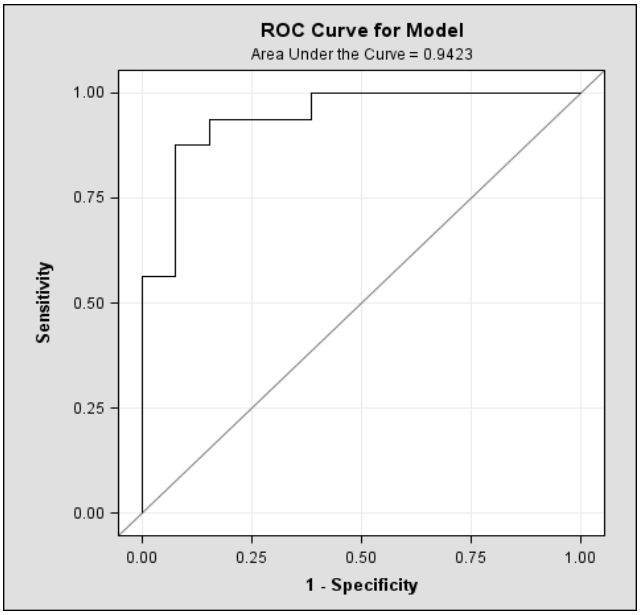


Table 3-14 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden Index
1	0.999	0.08	0.00	0.92	0.08
2	0.998	0.15	0.00	0.85	0.15
3	0.998	0.23	0.00	0.77	0.23
4	0.988	0.31	0.00	0.69	0.31
5	0.988	0.38	0.00	0.62	0.38
6	0.983	0.46	0.00	0.54	0.46
7	0.972	0.54	0.00	0.46	0.54
8	0.967	0.62	0.00	0.38	0.62
9	0.774	0.62	0.06	0.39	0.56
10	0.765	0.69	0.06	0.31	0.63
11	0.718	0.77	0.06	0.24	0.71
12	0.655	0.85	0.06	0.17	0.79
13	0.443	0.85	0.13	0.20	0.72
14	0.421	0.92	0.13	0.15	0.79
15	0.337	0.92	0.19	0.20	0.73
16	0.279	0.92	0.25	0.26	0.67
17	0.136	0.92	0.31	0.32	0.61
18	0.111	0.92	0.38	0.38	0.54
19	0.086	0.92	0.44	0.44	0.48
20	0.075	1.00	0.44	0.44	0.56
21	0.064	1.00	0.50	0.50	0.5
22	0.045	1.00	0.56	0.56	0.44
23	0.035	1.00	0.63	0.63	0.37
24	0.034	1.00	0.69	0.69	0.31
25	0.033	1.00	0.75	0.75	0.25
26	0.029	1.00	0.81	0.81	0.19
27	0.026	1.00	0.88	0.88	0.12
28	0.023	1.00	0.94	0.94	0.06
29	0.019	1.00	1.00	1.00	0.00

The relationship between the predicted probability of severe breast cancer and severity of breast cancer is shown in Figure 3-12. From Figure 3-12, it can be seen that if 0.421 is set as the threshold for the prediction of severe breast cancer, two patients with moderate breast cancer condition are predicted to have severe breast cancer, while one patient with severe breast cancer condition is predicted to have moderate breast cancer. Figure 3-13 shows the confidence interval for the probability of having breast cancer for each person. There is only one patient with severe breast cancer, which has a confidence interval below 0.421.

For the two patients who have moderate breast cancer condition but predicted to have severe breast cancer, the lower bounds of confidence intervals for probabilities of having severe breast cancer are smaller than 0.421. This figure is shown here to illustrate the uncertainty in estimated probabilities. It is not repeated elsewhere since small sample sizes used here produce wide intervals. Future analyses with large datasets will yield more useable interval estimates.

The sensitivity and specificity for this test are 0.92 and 0.87, respectively. Table 3-15 summarizes how many persons are predicted to have severe breast cancer and how many persons are predicted to have moderate breast cancer for patients at different stages.

Figure 3-12 Prediction of existence of cancer based on the optimal cut-off probability

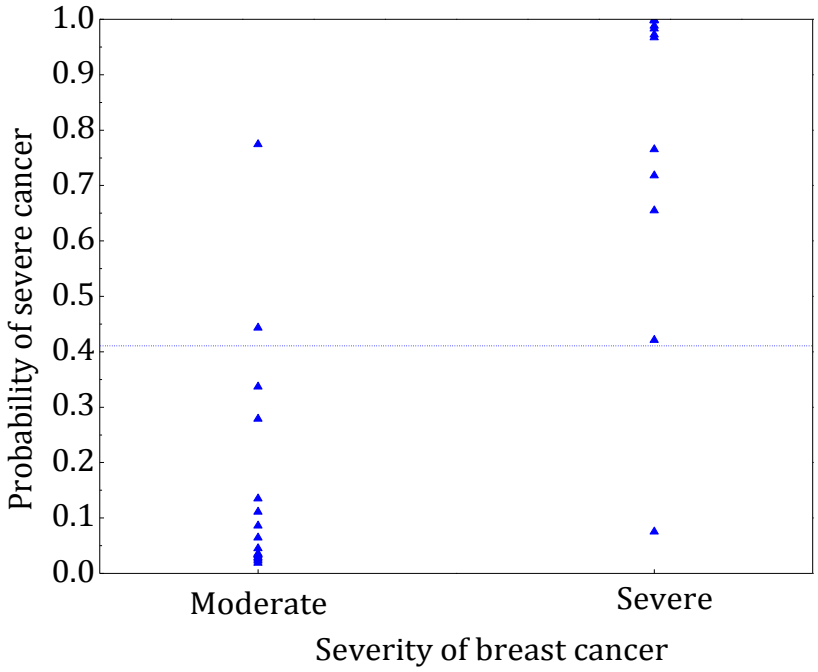
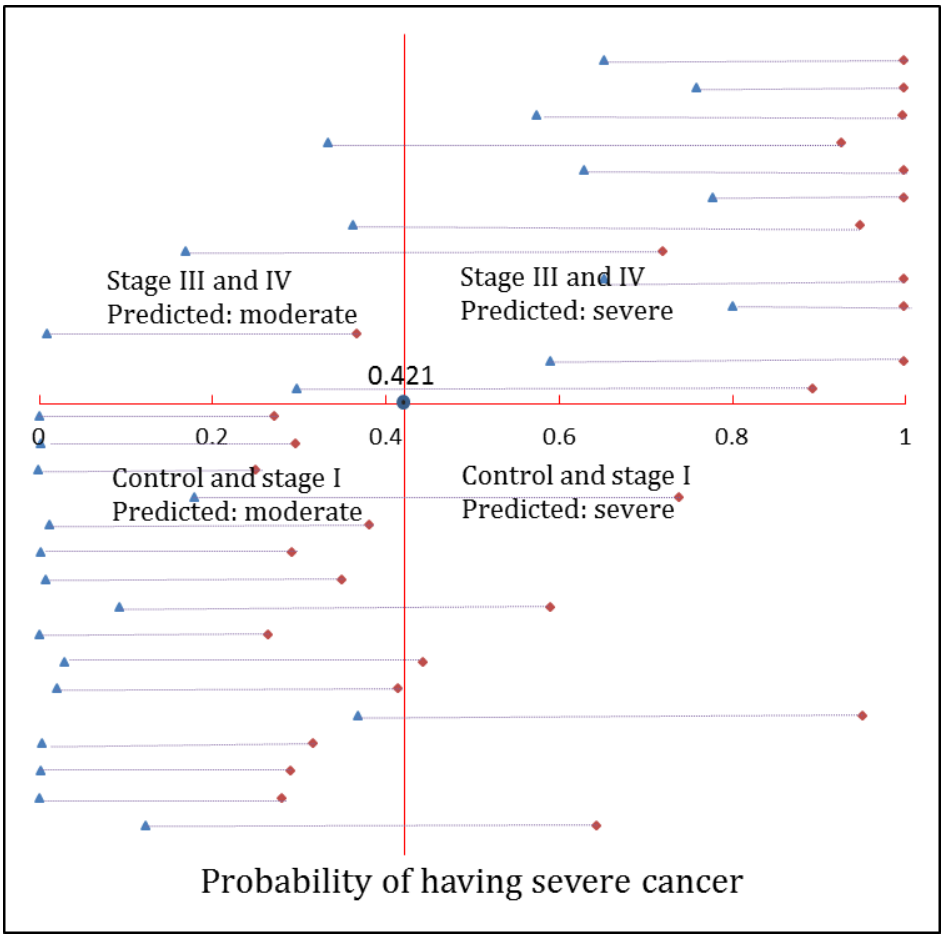


Table 3-15 Prediction of severity of breast cancer for patients in different stages

	Predicted	
	Moderate	Severe
C	11	1
0	1	1
I	2	0
III	1	5
IV	0	7

Figure 3-13 Confidence interval of probability of having breast cancer for each patient



3.4.2 Use of average CathB and age as the predictor

In this section, average CathB and age are used as the predictors to predict if the person has severe breast cancer. Estimates of parameters are shown in Table 3-16. Probability of having breast cancer for each patient is shown in Table 3-17.

Table 3-16 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.4669	4.0254	5.5309	0.0187
CathB	1	3.254E-6	1.189E-6	7.4885	0.0062
Age	1	0.0879	0.0564	2.4336	0.1188

Table 3-17 Probability of having breast cancer for each patient

Obs	Patient	Age	Staging	CathB	level	probability
1	B1	41	0	1353355	moderate	0.189
2	B2	64	IV	1770087	severe	0.872
3	B3	45	0	254411	moderate	0.009
4	B4	59	I	442882.7	moderate	0.055
5	B5	65	III	1611268	severe	0.816
6	B6	77	IV	3335243	severe	1.000
7	B6a	81	I	347228.3	moderate	0.228
8	B7	78	IV	2626708	severe	0.997
9	B8	75	III	2473030	severe	0.994
10	B11	46	III	677698.3	severe	0.038
11	B12	62	IV	1697287	severe	0.818
12	B13	40	IV	2419770	severe	0.872
13	B15	36	III	3481133	severe	0.993
14	B16	53	IV	3226621	severe	0.997
15	B17	52	III	2723205	severe	0.981
16	B18	43	IV	2720022	severe	0.959
17	B20	80	III	1327514	severe	0.868
18	C1	47	C	1220798	moderate	0.204
19	C2	51	C	1785704	moderate	0.696
20	C3	28	C	806983.7	moderate	0.012
21	C4	32	C	873253.3	moderate	0.022
22	C5	62	C	313112.3	moderate	0.048
23	C6	48	C	1140311	moderate	0.177
24	C7	49	C	627024.3	moderate	0.042
25	C8	26	C	428675.7	moderate	0.003
26	C9	42	C	720771.7	moderate	0.031
27	C10	60	C	387576	moderate	0.051
28	C11	26	C	420150.7	moderate	0.003
29	C12	44	C	515214.7	moderate	0.019

Table 3-18 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 3-14 is the ROC curve. From Table 3-18, it is found that 0.816 is the optimal cut-off probability to predict if the person has severe breast cancer. If the predicted probability of a person is above 0.816, this person is predicted to have severe breast cancer. If the predicted probability of a person is below 0.816, this person is predicted to have moderate breast cancer. The area under the ROC curve is 0.9567, indicating this test method is excellent.

Figure 3-14 ROC curve of the model

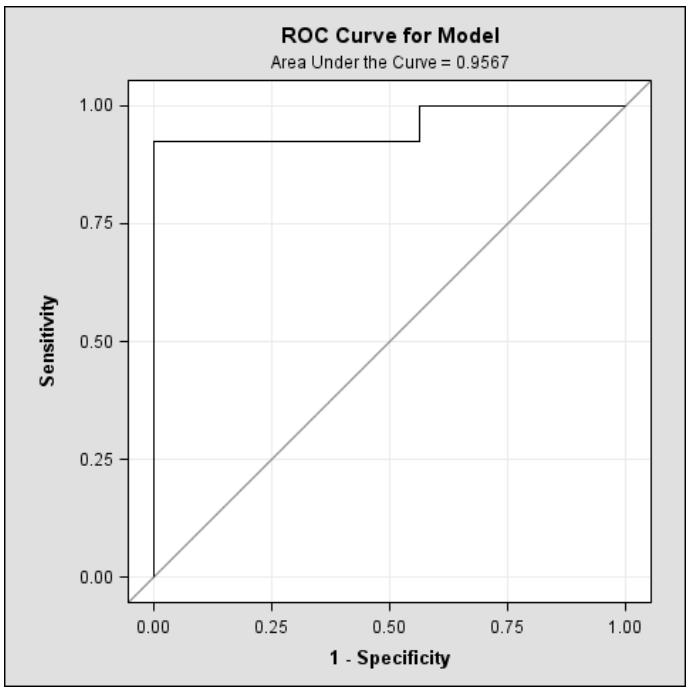


Table 3-18 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0,1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	1.000	0.08	0.00	0.92	0.08
2	0.997	0.15	0.00	0.85	0.15
3	0.997	0.23	0.00	0.77	0.23
4	0.994	0.31	0.00	0.69	0.31
5	0.993	0.38	0.00	0.62	0.38
6	0.981	0.46	0.00	0.54	0.46
7	0.959	0.54	0.00	0.46	0.54
8	0.873	0.62	0.00	0.38	0.62
9	0.872	0.69	0.00	0.31	0.69
10	0.868	0.77	0.00	0.23	0.77
11	0.819	0.85	0.00	0.15	0.85
12	0.816	0.92	0.00	0.08	0.92
13	0.696	0.92	0.06	0.10	0.86
14	0.229	0.92	0.13	0.15	0.79
15	0.204	0.92	0.19	0.20	0.73
16	0.189	0.92	0.25	0.26	0.67
17	0.177	0.92	0.31	0.32	0.61
18	0.055	0.92	0.38	0.38	0.54
19	0.051	0.92	0.44	0.44	0.48
20	0.048	0.92	0.50	0.51	0.42

21	0.042	0.92	0.56	0.57	0.36
22	0.039	1.00	0.56	0.56	0.44
23	0.031	1.00	0.63	0.63	0.37
24	0.022	1.00	0.69	0.69	0.31
25	0.019	1.00	0.75	0.75	0.25
26	0.012	1.00	0.81	0.81	0.19
27	0.009	1.00	0.88	0.88	0.12
28	0.003	1.00	0.94	0.94	0.06
29	0.003	1.00	1.00	1.00	0

When the cut-off probability is set to be 0.816, the relationship between the predicted existence of severe breast cancer and the actual diagnosis is shown in Figure 3-15. No patient with moderate breast cancer condition is predicted to have severe breast cancer, while one patient with severe breast cancer condition is predicted to have moderate breast cancer. The sensitivity and specificity for this test is 0. 92 and 1.00, respectively. Table 3-19 summarizes how many persons are predicted to have severe breast cancer and how many persons are predicted to have moderate breast cancer for patients at different stages.

Figure 3-15 Prediction of existence of cancer based on the optimal cut-off probability

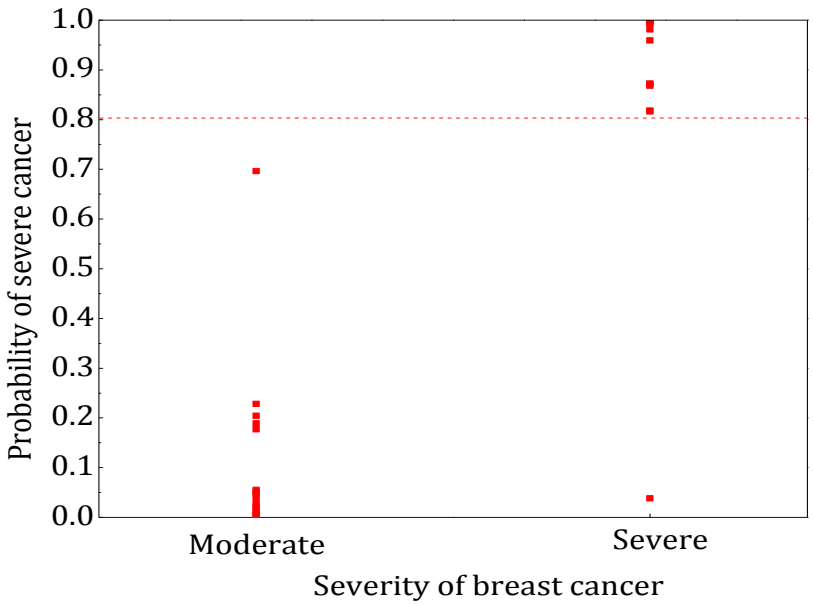


Table 3-19 Prediction of severity of breast cancer for patients at different stages of breast cancer

	Predicted	
	Moderate	Severe
C	12	0
0	2	0
I	2	0
III	1	5
IV	0	7

3.4.3 Comparison of models in section 3.4.1 and 3.4.2

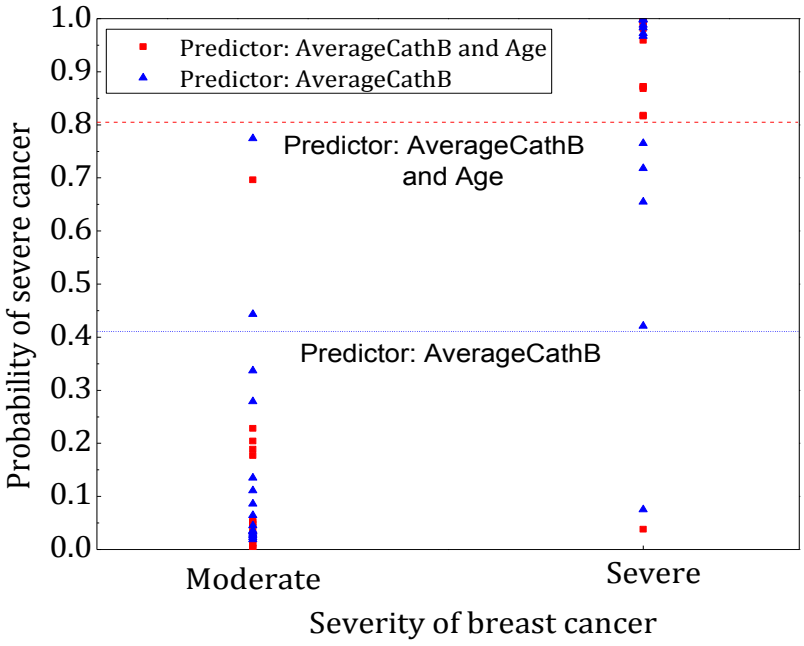
Probabilities computed in section 3.4.1 and section 3.4.2 are presented in Table 3-20. From Table 3-20, it can be seen that the probabilities calculated are quite similar for the persons whose ages are far from the average age of the persons involved in the test. But for the persons whose ages are close to the average age, the difference between the two probabilities is big. Comparison of the predicted results is shown in Figure 3-16. Including of age predictor increases the area under the ROC curve and increases sensitivity and specificity.

Table 3-20 Comparison of probabilities of having breast cancer for each patient between the two models in section 3.4.1 and 3.4.2

Obs	Patient	Staging	CathB	level	Prob1	Age	Prob2
1	B1	0	1353355	moderate	0.443	41	0.189
2	B11	III	677698.3	severe	0.075	46	0.038
3	B12	IV	1697287	severe	0.718	62	0.818
4	B13	IV	2419770	severe	0.967	40	0.872
5	B15	III	3481133	severe	0.999	36	0.993
6	B16	IV	3226621	severe	0.998	53	0.997
7	B17	III	2723205	severe	0.988	52	0.981
8	B18	IV	2720022	severe	0.988	43	0.959
9	B2	IV	1770087	severe	0.765	64	0.872
10	B20	III	1327514	severe	0.421	80	0.868
11	B3	0	254411	moderate	0.019	45	0.009
12	B4	I	442882.7	moderate	0.035	59	0.055
13	B5	III	1611268	severe	0.655	65	0.816
14	B6	IV	3335243	severe	0.998	77	1.000
15	B6a	I	347228.3	moderate	0.026	81	0.228
16	B7	IV	2626708	severe	0.983	78	0.997

17	B8	III	2473030	severe	0.972	75	0.994
18	C1	C	1220798	moderate	0.337	47	0.204
19	C10	C	387576	moderate	0.029	60	0.051
20	C11	C	420150.7	moderate	0.033	26	0.003
21	C12	C	515214.7	moderate	0.045	44	0.019
22	C2	C	1785704	moderate	0.774	51	0.696
23	C3	C	806983.7	moderate	0.111	28	0.012
24	C4	C	873253.3	moderate	0.135	32	0.022
25	C5	C	313112.3	moderate	0.023	62	0.048
26	C6	C	1140311	moderate	0.279	48	0.177
27	C7	C	627024.3	moderate	0.064	49	0.042
28	C8	C	428675.7	moderate	0.034	26	0.003
29	C9	C	720771.7	moderate	0.086	42	0.031

Figure 3-16 Prediction of existence of cancer based on the optimal cut-off probability for the models used in section 3.4.1 and 3.4.2



3.5 Analysis based on logistic regression with binary response (complete data)

As one last grouping, patients at stage III and IV are group together and their breast cancer condition is severe. All the other individuals belong to the control group, and their breast cancer condition is denoted as moderate. The reason for division of patients into such two groups is explained in section 3.2, and is done here to make use of all samples.

3.5.1 Use of average CathB is used as the predictor

When average CathB is used as the predictor, the estimates of parameters are shown in Table 3-21. The predicted probability of severe breast cancer for each person is listed in Table 3-22.

Table 3-21 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Error	Standard Chi-Square	Wald Pr > ChiSq
Intercept	1	-5.2329	1.8366	8.1178	0.0044
CathB	1	3.535E-6	1.308E-6	7.3004	0.0069

Table 3-22 Probability of having breast cancer for each patient

Obs	Patient	Staging	CathB	Level	Prob
1	C1	C	1220798	moderate	0.285
2	C2	C	1785704	moderate	0.746
3	C3	C	806983.7	moderate	0.085
4	C4	C	873253.3	moderate	0.105
5	C5	C	313112.3	moderate	0.016
6	C6	C	1140311	moderate	0.231
7	C7	C	627024.3	moderate	0.047
8	C8	C	428675.7	moderate	0.024
9	C9	C	720771.7	moderate	0.064
10	C10	C	387576	moderate	0.021
11	C11	C	420150.7	moderate	0.023
12	C12	C	515214.7	moderate	0.032
13	B1	0	1353355	moderate	0.390
14	B3	0	254411	moderate	0.013
15	B4	I	442882.7	moderate	0.025
16	B6a	I	347228.3	moderate	0.018
17	B10	II	1264716	moderate	0.318
18	B14	II	1060055	moderate	0.185
19	B19	II	456155	moderate	0.026
20	B5	III	1611268	severe	0.614
21	B8	III	2473030	severe	0.971
22	B11	III	677698.3	severe	0.055
23	B15	III	3481133	severe	0.999
24	B17	III	2723205	severe	0.988
25	B20	III	1327514	severe	0.368
26	B2	IV	1770087	severe	0.736
27	B6	IV	3335243	severe	0.999
28	B7	IV	2626708	severe	0.983

29	B12	IV	1697287	severe	0.683
30	B13	IV	2419770	severe	0.965
31	B16	IV	3226621	severe	0.998
32	B18	IV	2720022	severe	0.988

The optimal cut-off probability to predict if the person has severe breast cancer is found by a ROC curve. Table 3-23 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 3-17 is the ROC curve. The optimal point is, again the one which has the smallest distance to the point (0, 1) and at the same time has the largest vertical distance to the line of equality. Based on these criteria, the probability of 0.368 is the optimal cut-off probability to predict if the person has severe breast cancer. If the predicted probability of a person is above 0.368, this person is predicted to have severe breast cancer. If the predicted probability of a person is below 0.368, this person is predicted to have moderate breast cancer. The area under the ROC curve is 0.9433, indicating this test method is excellent.

Figure 3-17 ROC curve of the model

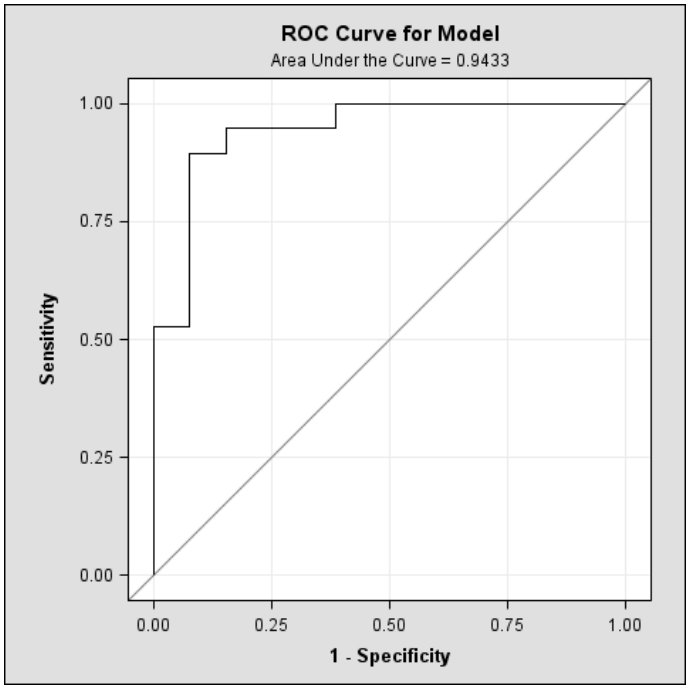


Table 3-23 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	0.999	0.08	0.00	0.92	0.08
2	0.999	0.15	0.00	0.85	0.15
3	0.998	0.23	0.00	0.77	0.23
4	0.988	0.31	0.00	0.69	0.31
5	0.988	0.38	0.00	0.62	0.38
6	0.983	0.46	0.00	0.54	0.46
7	0.971	0.54	0.00	0.46	0.54
8	0.965	0.62	0.00	0.38	0.62
9	0.747	0.62	0.05	0.39	0.57
10	0.736	0.69	0.05	0.31	0.64
11	0.683	0.77	0.05	0.24	0.72
12	0.614	0.85	0.05	0.16	0.8
13	0.390	0.85	0.11	0.19	0.74
14	0.368	0.92	0.11	0.13	0.81
15	0.318	0.92	0.16	0.18	0.76
16	0.286	0.92	0.21	0.22	0.71
17	0.231	0.92	0.26	0.27	0.66
18	0.185	0.92	0.32	0.33	0.6
19	0.105	0.92	0.37	0.38	0.55
20	0.085	0.92	0.42	0.43	0.5
21	0.064	0.92	0.47	0.48	0.45
22	0.055	1.00	0.47	0.47	0.53
23	0.047	1.00	0.53	0.53	0.47
24	0.032	1.00	0.58	0.58	0.42
25	0.026	1.00	0.63	0.63	0.37
26	0.025	1.00	0.68	0.68	0.32
27	0.024	1.00	0.74	0.74	0.26
28	0.023	1.00	0.79	0.79	0.21
29	0.021	1.00	0.84	0.84	0.16
30	0.018	1.00	0.89	0.89	0.11
31	0.016	1.00	0.95	0.95	0.05
32	0.013	1.00	1.00	1.00	0

Relationship between the predicted probability of severe breast cancer and severity of breast cancer is shown in Figure 3-18. From Figure 3-18, it can be seen that if 0.368 is set as the threshold for the prediction of severe breast cancer, two patients with moderate breast cancer condition are predicted to have severe breast cancer, while one patient with severe breast cancer condition is predicted to have moderate breast cancer.

The sensitivity and specificity for this test is 0.92 and 0.89, respectively. Table 3-24 summarizes how many persons are predicted to have severe breast cancer and how many persons are predicted to have moderate breast cancer for patients at different stages.

Figure 3-18 Prediction of existence of cancer based on the optimal cut-off probability

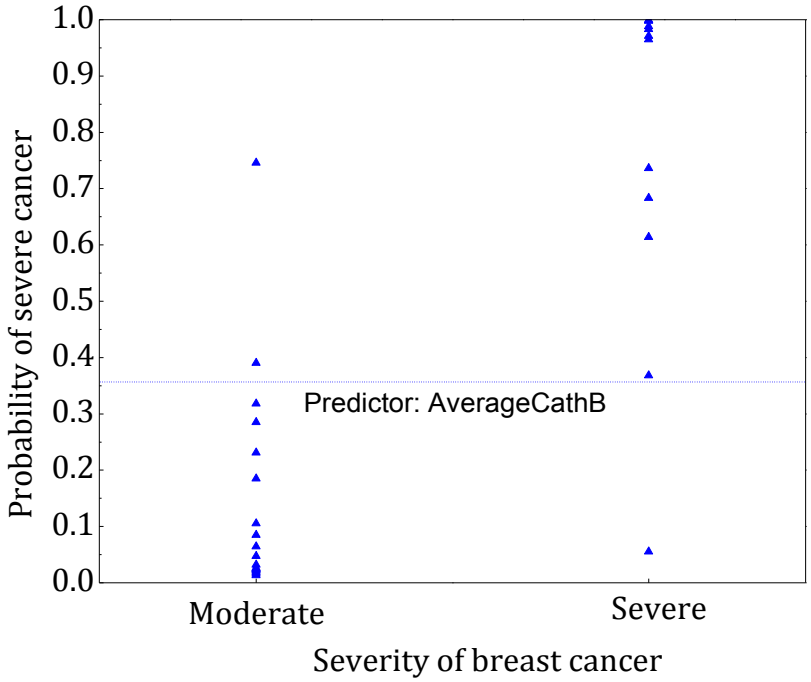


Table 3-24 Prediction of severity of breast cancer for patients at different stages

	Predicted	
	Moderate	Severe
C	11	1
0	1	1
I	2	0
II	3	0
III	1	5
IV	0	7

3.5.2 Use of average CathB and Age used as the predictor

In this section, average CathB and age are used as the predictors to predict if the person has severe breast cancer. Estimates of parameters are shown in Table 3-25. Probability of having breast cancer for each patient is shown in Table 3-26.

Table 3-25 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-9.9959	4.1125	5.9078	0.0151
CathB	1	3.374E-6	1.214E-6	7.7236	0.0055
Age	1	0.0898	0.0568	2.4994	0.1139

Table 3-26 Probability of having breast cancer for each patient

Obs	Patient	Age	Staging	CathB	level	Prob
1	B1	41	0	1353355	moderate	0.148
2	B2	64	IV	1770087	severe	0.849
3	B3	45	0	254411	moderate	0.006
4	B4	59	I	442882.7	moderate	0.039
5	B5	65	III	1611268	severe	0.782
6	B6	77	IV	3335243	severe	1.000
7	B6a	81	I	347228.3	moderate	0.175
8	B7	78	IV	2626708	severe	0.997
9	B8	75	III	2473030	severe	0.994
10	B10	59	II	1264716	moderate	0.394
11	B11	46	III	677698.3	severe	0.027
12	B12	62	IV	1697287	severe	0.786
13	B13	40	IV	2419770	severe	0.853
14	B14	40	II	1060055	moderate	0.056
15	B15	36	III	3481133	severe	0.993
16	B16	53	IV	3226621	severe	0.997
17	B17	52	III	2723205	severe	0.979
18	B18	43	IV	2720022	severe	0.954
19	B19	51	II	456155	moderate	0.020
20	B20	80	III	1327514	severe	0.841
21	C1	47	C	1220798	moderate	0.160
22	C2	51	C	1785704	moderate	0.648
23	C3	28	C	806983.7	moderate	0.009
24	C4	32	C	873253.3	moderate	0.015
25	C5	62	C	313112.3	moderate	0.033
26	C6	48	C	1140311	moderate	0.137
27	C7	49	C	627024.3	moderate	0.030
28	C8	26	C	428675.7	moderate	0.002
29	C9	42	C	720771.7	moderate	0.022
30	C10	60	C	387576	moderate	0.036
31	C11	26	C	420150.7	moderate	0.002
32	C12	44	C	515214.7	moderate	0.013

Table 3-27 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 3-19 is the ROC curve. From Table 3-26, it is found that 0.782 is the optimal cut-off probability to predict if the person has severe breast cancer. If the predicted probability of a person is above 0.782, this person is predicted to have severe breast cancer. If the predicted probability of a person is below 0.782, this person is predicted to have moderate breast cancer. The area under the ROC curve is 0.9555, indicating this test method is excellent.

Figure 3-19 ROC curve of the model

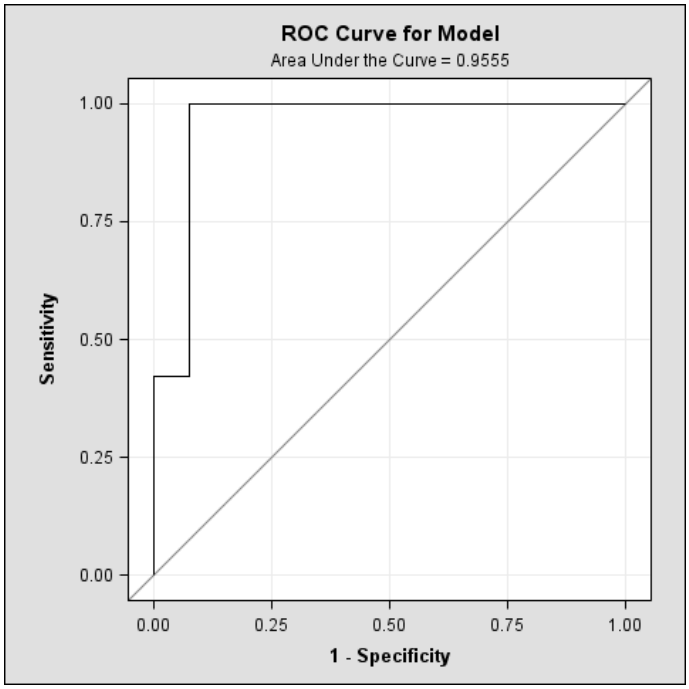


Table 3-27 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	1.000	0.08	0.00	0.92	0.08
2	0.997	0.15	0.00	0.85	0.15
3	0.997	0.23	0.00	0.77	0.23
4	0.994	0.31	0.00	0.69	0.31
5	0.993	0.38	0.00	0.62	0.38
6	0.979	0.46	0.00	0.54	0.46
7	0.955	0.54	0.00	0.46	0.54
8	0.853	0.62	0.00	0.38	0.62
9	0.849	0.69	0.00	0.31	0.69

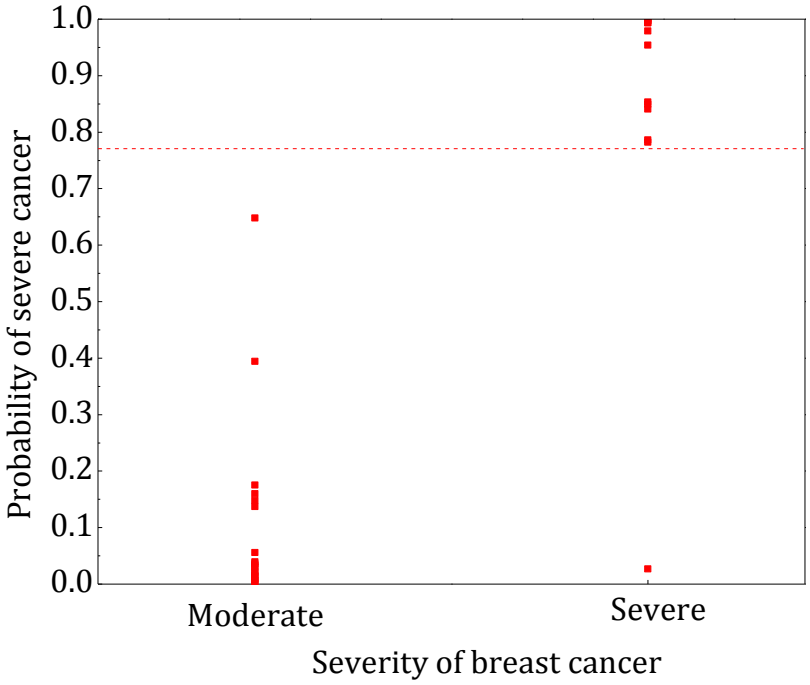
10	0.841	0.77	0.00	0.23	0.77
11	0.786	0.85	0.00	0.15	0.85
12	0.782	0.92	0.00	0.08	0.92
13	0.648	0.92	0.05	0.09	0.87
14	0.394	0.92	0.11	0.13	0.81
15	0.175	0.92	0.16	0.18	0.76
16	0.160	0.92	0.21	0.22	0.71
17	0.148	0.92	0.26	0.27	0.66
18	0.137	0.92	0.32	0.33	0.6
19	0.056	0.92	0.37	0.38	0.55
20	0.039	0.92	0.42	0.43	0.5
21	0.036	0.92	0.47	0.48	0.45
22	0.033	0.92	0.53	0.53	0.39
23	0.030	0.92	0.58	0.58	0.34
24	0.027	1.00	0.58	0.58	0.42
25	0.022	1.00	0.63	0.63	0.37
26	0.020	1.00	0.68	0.68	0.32
27	0.015	1.00	0.74	0.74	0.26
28	0.013	1.00	0.79	0.79	0.21
29	0.009	1.00	0.84	0.84	0.16
30	0.006	1.00	0.89	0.89	0.11
31	0.002	1.00	0.95	0.95	0.05
32	0.002	1.00	1.00	1.00	0

When the cut-off probability is set to be 0.782, the relationship between the predicted existence of severe breast cancer and the actual diagnosis is shown in Figure 3-20. No patient with moderate breast cancer condition is predicted to have severe breast cancer, while one patient with severe breast cancer condition is predicted to have moderate breast cancer. The sensitivity and specificity for this test is 0.92 and 1.00, respectively. Table 3-28 summarizes how many persons are predicted to have severe breast cancer and how many persons are predicted to have moderate breast cancer for patients at different stages.

Table 3-28 Prediction of severity of breast cancer for patients at different stages

	Predicted	
	Moderate	Severe
C	12	0
0	2	0
I	2	0
II	3	0
III	1	5
IV	0	7

Figure 3-20 Prediction of existence of cancer based on the optimal cut-off probability



3.5.3 Comparison of models in section 3.5.1 and 3.5.2

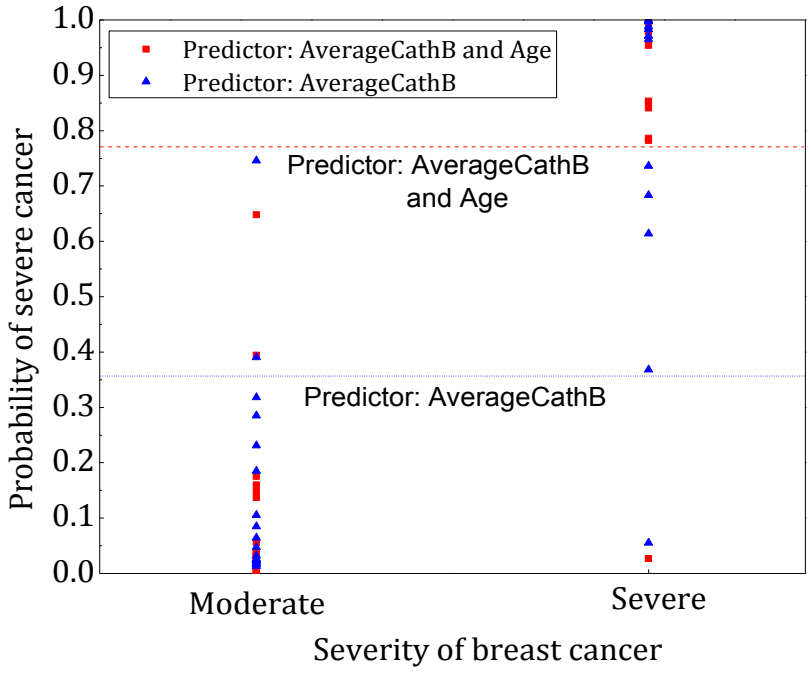
Probabilities computed in section 3.5.1 and section 3.5.2 are presented in Table 3-29. From Table 3-28, it can be seen that the probabilities calculated are quite similar for patients whose ages are close to the average age of the group, indicating age plays a role in the analysis. Comparison of the predicted results is shown in Figure 3-21. Including of age predictor increases the area under the ROC curve and increases sensitivity and prediction.

Table 3-29 Comparison of probabilities of having breast cancer for each patient between the two models in section 3.5.1 and 3.5.2

Obs	Patient	Staging	CathB	level	Prob1	Age	Prob2
1	B1	0	1353355	moderate	0.390	41	0.148
2	B10	II	1264716	moderate	0.318	59	0.394
3	B11	III	677698.3	severe	0.055	46	0.027
4	B12	IV	1697287	severe	0.683	62	0.786
5	B13	IV	2419770	severe	0.965	40	0.853
6	B14	II	1060055	moderate	0.185	40	0.056
7	B15	III	3481133	severe	0.999	36	0.993
8	B16	IV	3226621	severe	0.998	53	0.997
9	B17	III	2723205	severe	0.988	52	0.979
10	B18	IV	2720022	severe	0.988	43	0.954

11	B19	II	456155	moderate	0.026	51	0.020
12	B2	IV	1770087	severe	0.736	64	0.849
13	B20	III	1327514	severe	0.368	80	0.841
14	B3	0	254411	moderate	0.013	45	0.006
15	B4	I	442882.7	moderate	0.025	59	0.039
16	B5	III	1611268	severe	0.614	65	0.782
17	B6	IV	3335243	severe	0.999	77	1.000
18	B6a	I	347228.3	moderate	0.018	81	0.175
19	B7	IV	2626708	severe	0.983	78	0.997
20	B8	III	2473030	severe	0.971	75	0.994
21	C1	C	1220798	moderate	0.285	47	0.160
22	C10	C	387576	moderate	0.021	60	0.036
23	C11	C	420150.7	moderate	0.023	26	0.002
24	C12	C	515214.7	moderate	0.032	44	0.013
25	C2	C	1785704	moderate	0.746	51	0.648
26	C3	C	806983.7	moderate	0.085	28	0.009
27	C4	C	873253.3	moderate	0.105	32	0.015
28	C5	C	313112.3	moderate	0.016	62	0.033
29	C6	C	1140311	moderate	0.231	48	0.137
30	C7	C	627024.3	moderate	0.047	49	0.030
31	C8	C	428675.7	moderate	0.024	26	0.002
32	C9	C	720771.7	moderate	0.064	42	0.022

Figure 3-21 Prediction of existence of cancer based on the optimal cut-off probability for the models used in section 3.5.1 and 3.5.2



3.6 Importance of age variable

Table 3-30 is a summary of the importance of the age variable. For all the three datasets, including age as one the predictor increases the area under the ROC curve, and increases optimal cut-off probability. Except for the results in section 3.3 where stages 0 and I are deleted, including age variable also increases sensitivity and specificity. As a result, it is recommended that for similar research situation, age should be used as a predictor in the model.

Table 3-30 Effects of age factor and deletion of stages on area under the ROC curve, optimal cut-off probability, sensitivity, specificity, and p-value

Predictor	Division of two groups	Area under the ROC	Cut-off Prob	Sensi	Speci	P-value	
						CathB	Age
Average CathB	Cancer (II, III, IV) No cancer (C)	0.8854	0.58	0.81	0.92	0.0102	N/A
	Moderate cancer (C, 0, I) Severe cancer (III, IV)	0.9423	0.421	0.92	0.87	0.0072	N/A
	Moderate cancer (C, 0, I, II) Severe cancer (III, IV)	0.9433	0.368	0.92	0.89	0.0069	N/A
Average CathB+Age	Cancer (II, III, IV) No cancer (C)	0.9063	0.723	0.81	0.92	0.0177	0.1431
	Moderate cancer (C, 0, I) Severe cancer (III, IV)	0.9567	0.816	0.92	1	0.0062	0.1188
	Moderate cancer (C, 0, I, II) Severe cancer (III, IV)	0.9555	0.782	0.92	1	0.0055	0.1139

3.7 Analysis based on multicategory logistic model (complete data)

In this section, age and average CathB are used as the predictors to predict the probabilities of each stage of cancer for different persons. Breast cancer of staging 0 and control group are combined as the control group. There are in total five responses: control, staging I, staging II, staging III and staging IV. Analysis of effects of different parameters is shown in Table 3-31. Estimates of parameters are shown in Table 3-32. Table 3-32 lists the intercepts and coefficients of the four equations fit for the four cancer stages, I, II, III, and IV. The intercept and coefficient of the equation for control group are set to be 0 by default. Probabilities of each stage of breast cancer for different persons are shown in Table 3-33. For example, the first line of Table 3-33, for the first observation, the probability that this

person does not have breast cancer or has a breast cancer at stage 0 is 0.727, whereas, the probabilities of have breast cancer at stage I, II, III, or IV is 0.000, 0.149, 0.090, and 0.034, respectively. The p-values of average CathB and age are 0.0771 and 0.3384, indicating average CathB has a marginal effect on stage of breast cancer, whereas age has little effect on stage of breast cancer.

Table 3-31 Analysis of Effects

Effect	DF	Chi-Square	Pr > ChiSq
average CathB	4	8.4291	0.0771
Age	4	4.5348	0.3384

Table 3-32 Coefficients and intercept of the four logistic regression equations fit for the four cancer stages

Parameter	Staging	DF	Estimate	Error	Standard Chi-square	Wald Pr>ChiSq
Intercept	IV	1	-14.1369	5.4963	6.6155	0.0101
Intercept	III	1	-11.3583	4.7944	5.6126	0.0178
Intercept	II	1	-4.7017	3.2863	2.0469	0.1525
Intercept	I	1	-12.3027	8.295	2.1997	0.138
average CathB	IV	1	3.85E-06	1.43E-06	7.2307	0.0072
average CathB	III	1	2.92E-06	1.28E-06	5.2589	0.0218
average CathB	II	1	5.45E-07	1.25E-06	0.1904	0.6626
average CathB	I	1	-3.33E-06	4.46E-06	0.5602	0.4542
Age	IV	1	0.143	0.0806	3.1516	0.0759
Age	III	1	0.1295	0.0763	2.8835	0.0895
Age	II	1	0.0581	0.0625	0.8627	0.353
Age	I	1	0.2163	0.1405	2.3676	0.1239

For each person, the stage of cancer is predicted to be the one with the largest probability. As a summary, table 3-34 lists the numbers of persons predicted to have breast cancers of different stages.

Table 3-33 Probabilities of breast cancer in each stage for each patient

Obs	Age	Staging	CathB	Prob at stage 0 or no cancer	Prob at stage I	Prob at stage II	Prob at stage III	Prob at stage IV
1	41	C	1353355	0.727	0.000	0.149	0.090	0.034
2	64	IV	1770087	0.061	0.001	0.060	0.499	0.379
3	45	C	254411	0.844	0.028	0.120	0.007	0.001

4	59	I	442882.7	0.548	0.198	0.195	0.049	0.010
5	65	III	1611268	0.085	0.002	0.081	0.499	0.332
6	77	IV	3335243	0.000	0.000	0.000	0.206	0.794
7	81	I	347228.3	0.016	0.940	0.020	0.019	0.005
8	78	IV	2626708	0.001	0.000	0.002	0.330	0.668
9	75	III	2473030	0.001	0.000	0.004	0.370	0.625
10	59	II	1264716	0.334	0.008	0.186	0.327	0.145
11	46	III	677698.3	0.807	0.008	0.153	0.026	0.006
12	62	IV	1697287	0.095	0.001	0.080	0.487	0.337
13	40	IV	2419770	0.160	0.000	0.056	0.392	0.393
14	40	II	1060055	0.816	0.001	0.135	0.038	0.011
15	36	III	3481133	0.009	0.000	0.004	0.279	0.708
16	53	IV	3226621	0.002	0.000	0.002	0.283	0.713
17	52	III	2723205	0.014	0.000	0.011	0.381	0.594
18	43	IV	2720022	0.045	0.000	0.022	0.392	0.541
19	51	II	456155	0.755	0.046	0.170	0.025	0.005
20	80	III	1327514	0.030	0.052	0.058	0.529	0.331
21	47	C	1220798	0.657	0.001	0.178	0.120	0.043
22	51	C	1785704	0.245	0.000	0.114	0.390	0.252
23	28	C	806983.7	0.928	0.000	0.067	0.004	0.001
24	32	C	873253.3	0.905	0.000	0.085	0.009	0.002
25	62	C	313112.3	0.390	0.415	0.154	0.035	0.007
26	48	C	1140311	0.668	0.002	0.184	0.109	0.037
27	49	C	627024.3	0.773	0.017	0.170	0.032	0.007
28	26	C	428675.7	0.949	0.000	0.049	0.001	0.000
29	42	C	720771.7	0.844	0.003	0.130	0.019	0.004
30	60	C	387576	0.498	0.268	0.182	0.043	0.009
31	26	C	420150.7	0.949	0.000	0.049	0.001	0.000
32	44	C	515214.7	0.844	0.009	0.131	0.013	0.002

Table 3-34 Prediction of staging of breast cancer for patients at different stages of breast cancer

Staging	Predicted Stage				
	C	I	III	IV	Sum
C	10	1	1		12
I	1	1			2
II	3				3
III	1		2	3	6
IV			2	5	7
Sum	17	2	5	8	32

3.8 Analysis based on CART

The data are also analyzed by a classification tree. The aim is to predict if the person has severe or moderate breast cancer based on predictors of average CathB and age. The persons are divided into two groups using the same method as in section 3.5. Patients with breast cancer of stages III and IV are grouped together as severe breast cancer group. The others are grouped together as moderate breast cancer group. The output is shown in Figure 22. If the average CathB is greater than 1.296E06, the patient is predicted to have severe breast cancer. On the contrary, if the average CathB of the individual is smaller than 1.296E06, the individual is predicted to have moderate breast cancer. 13 persons are predicted to have severe breast cancer, whereas, 1 of them actually has moderate breast cancer. 19 persons are predicted to have moderate breast cancer, whereas, 2 of them have severe breast cancer. Table 3-35 shows the predicted breast cancer condition for each patient.

Figure 3-22 Prediction of breast cancer condition based on CART model

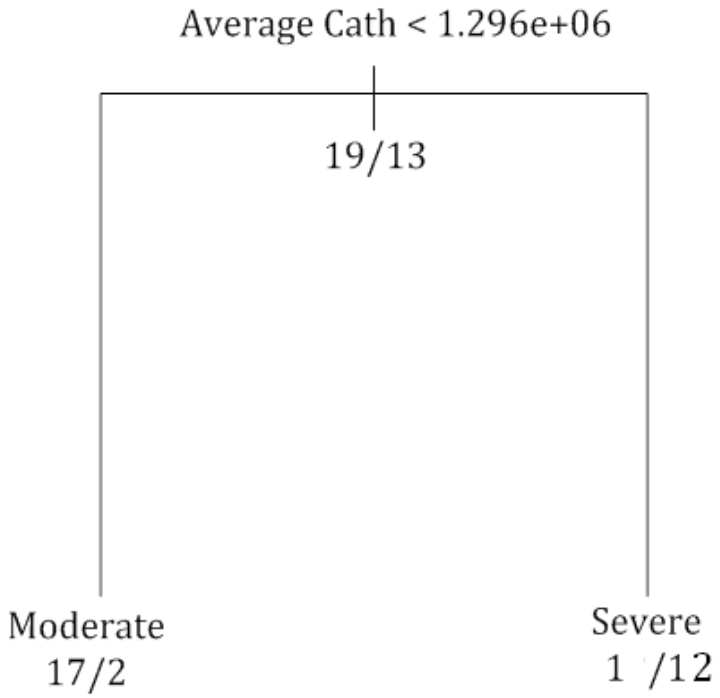
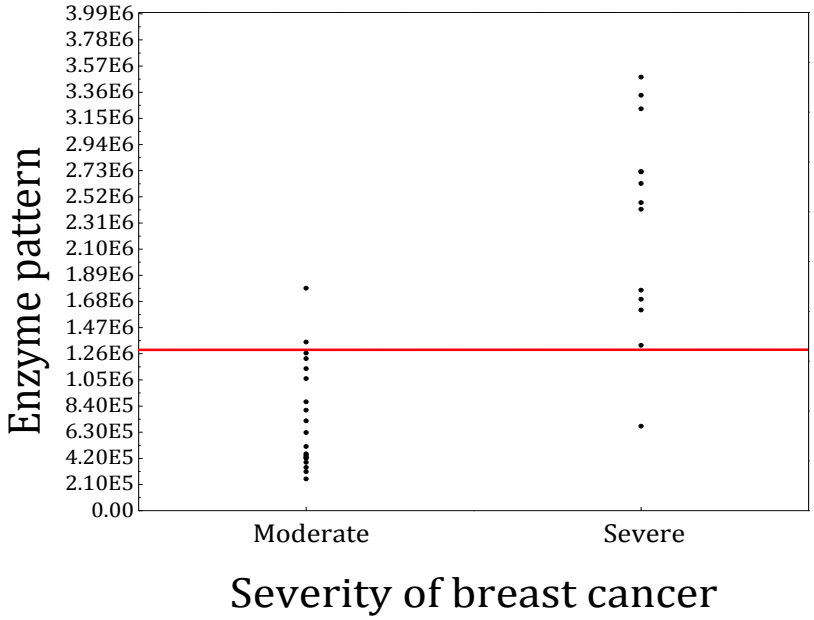


Table 3-35 Predicted breast cancer condition by tree model

Patient	Age	Staging	Level	averCathB	predicted
B1	41	0	moderate	1353355	Severe
B2	64	IV	severe	1770087	Severe
B3	45	0	moderate	254411	Moderate
B4	59	I	moderate	442882.7	Moderate
B5	65	III	severe	1611268	Severe
B6	77	IV	severe	3335243	Severe
B6a	81	I	moderate	347228.3	Moderate
B7	78	IV	severe	2626708	Severe
B8	75	III	severe	2473030	Severe
B10	59	II	moderate	1264716	Moderate
B11	46	III	severe	677698.3	Moderate
B12	62	IV	severe	1697287	Severe
B13	40	IV	severe	2419770	Severe
B14	40	II	moderate	1060055	Moderate
B15	36	III	severe	3481133	Severe
B16	53	IV	severe	3226621	Severe
B17	52	III	severe	2723205	Severe
B18	43	IV	severe	2720022	Severe
B19	51	II	moderate	456155	Moderate
B20	80	III	severe	1327514	Severe
C1	47	C	moderate	1220798	Moderate
C2	51	C	moderate	1785704	Severe
C3	28	C	moderate	806983.7	Moderate
C4	32	C	moderate	873253.3	Moderate
C5	62	C	moderate	313112.3	Moderate
C6	48	C	moderate	1140311	Moderate
C7	49	C	moderate	627024.3	Moderate
C8	26	C	moderate	428675.7	Moderate
C9	42	C	moderate	720771.7	Moderate
C10	60	C	moderate	387576	Moderate
C11	26	C	moderate	420150.7	Moderate
C12	44	C	moderate	515214.7	Moderate

Figure 3-23 Scatter plot of enzyme pattern vs severity of breast cancer



The prediction for patients in the moderate breast cancer group and in the severe breast cancer group is illustrated in Figure 3-23. It can be seen that two persons in the moderate breast cancer group are predicted to have severe breast cancer, whereas, one patient with severe breast cancer is predicted to have moderate breast cancer.

The measures to assess a diagnostic test are as follows:

(1) Sensitivity is the proportion of the patients with severe breast cancer diagnosis that are correctly identified by the test.

$$\text{Sensitivity} = \frac{number\ of\ patients\ with\ severe\ breast\ cancer\ that\ are\ correctly\ identified}{number\ of\ patients\ with\ severe\ breast\ cancer}$$

$$= \frac{12}{13} = 0.923$$

(2) Specificity is the proportion of the patients with moderate breast cancer diagnosis that are correctly identified by the test.

$$\text{Specificity} = \frac{number\ of\ patients\ with\ moderate\ breast\ cancer\ that\ are\ correctly\ identified}{number\ of\ patients\ with\ moderate\ breast\ cancer}$$

$$= \frac{17}{19} = 0.895$$

(3) Proportion of correct diagnoses for moderate patients is $\frac{17}{18} = 0.944$

(4) Proportion of corrected diagnoses for severe patients is $\frac{12}{14} = 0.857$

(5) Prevalence of abnormality is the proportion of severe patients among all the individuals in the test.

$$\text{Prevalence of abnormality} = \frac{number\ of\ severe\ patients}{number\ of\ individuals\ in\ the\ test} = \frac{13}{32} = 0.406$$

(6) Positive predictive value (PPV) is the proportion of patients with severe test results who are correctly diagnosed.

$$\text{Positive predictive value} = \frac{sensitivity * prevalence}{sensitivity * prevalence + (1 - specificity) * (1 - prevalence)}$$
$$= \frac{0.923 * 0.406}{0.923 * 0.406 + (1 - 0.895) * (1 - 0.406)} = 0.857$$

(7) Negative predictive value (NPV) is the proportion of patients with moderate test results who are correctly diagnosed.

$$\text{Negative predictive value} = \frac{specificity * prevalence}{(1 - sensitivity) * prevalence + specificity * (1 - prevalence)}$$
$$= \frac{0.895 * 0.406}{(1 - 0.923) * 0.406 + 0.895 * (1 - 0.406)} = 0.646$$

(8) Likelihood ratio is the ratio of the probability of getting a severe result if the patient truly had the severe breast cancer and the probability of getting severe results if the individual was healthy.

$$\text{Likelihood ratio} = \frac{sensitivity}{(1 - specificity)} = \frac{0.923}{1 - 0.895} = 8.79$$

It needs to note that the number of wrongly predicted patients is the same as the number obtained by the logistic regression model when average CathB is used as the predictor (model in section 3.5.1) and this grouping is used.

Chapter 4 - Assessing Diagnostic Test for Breast Cancer (data MMP1)

4.1 Overview of the data

Another type of nanoparticle is used to detect breast cancer by examining the enzyme pattern in the blood of a patient. The nanoparticles are coated with matrix metalloproteinase-1, which is a type of proteases to interact with the enzyme in the blood samples. Information on enzyme activity, patients' ages and stages of cancer were recorded. Relationships between average MMP, staging of cancer, and age of patient are shown in Figure 4-1and Figure 4-2. From these boxplots, it can be seen that the patients with cancer stages of I, II, III, and IV have higher average MMP than the persons in the control group, which includes healthy persons, and patients with a cancer stage of 0. But the differences among average MMP of patients with breast cancer of stage of I, II, III, and IV are not remarkable. Age of patient does not show a significant relationship with average MMP. As seen in chapter 3, age of patient does seem to have a relationship with staging of breast cancer. Patients with breast cancer of stage I, II, III, and IV are older than the persons without breast cancer or with breast cancer of stage 0.

Figure 4-1 Boxplot of average MMP vs staging of cancer

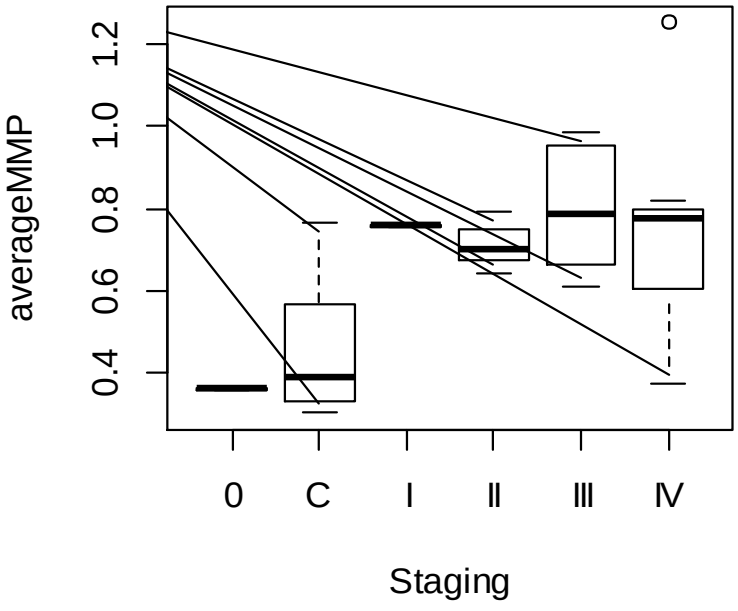
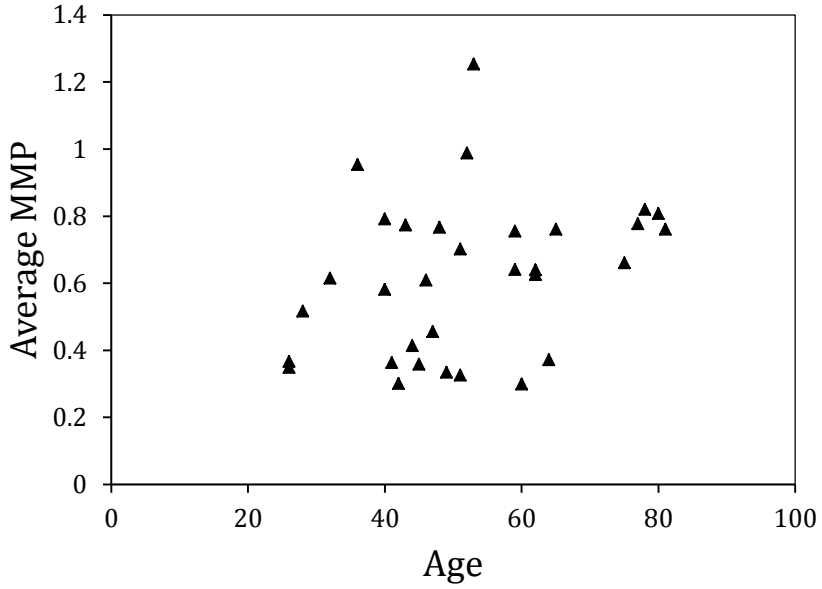


Figure 4-2 Scatterplot of average MMP vs age of patient



4.2 Comparison of three measurements

The enzyme pattern of each person was tested three times on different days. Figure 4-3 shows the results of the three measurements. It can be seen that the experimental error is very small. In addition, a logistic regression model is also used to predict the probability of having breast cancer for each person by using the individual enzyme pattern instead of the average enzyme pattern over the three measurements. If the person has breast cancer of stage I, II, III and IV, this person belongs to the "having breast cancer" group. If the person has no breast cancer or if he has breast cancer of stage 0, the person belongs to the "having no breast cancer" group. The probability of having breast cancer for each patient is shown in Figure 4-4, which shows that there is no noticeable difference between the three probabilities for all the persons. As a result, in this chapter, the models are built based on the average value of these three measurements.

Figure 4-3 Scatter plot of the three enzyme patterns of different patients

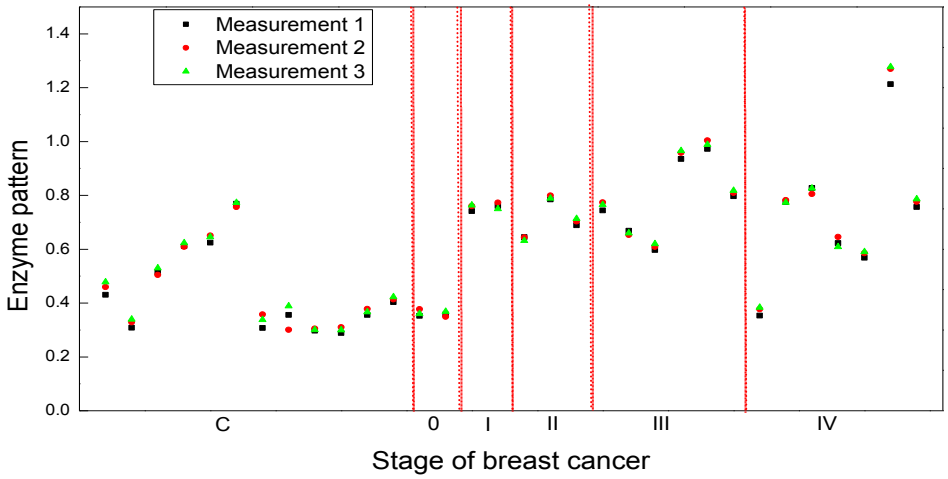
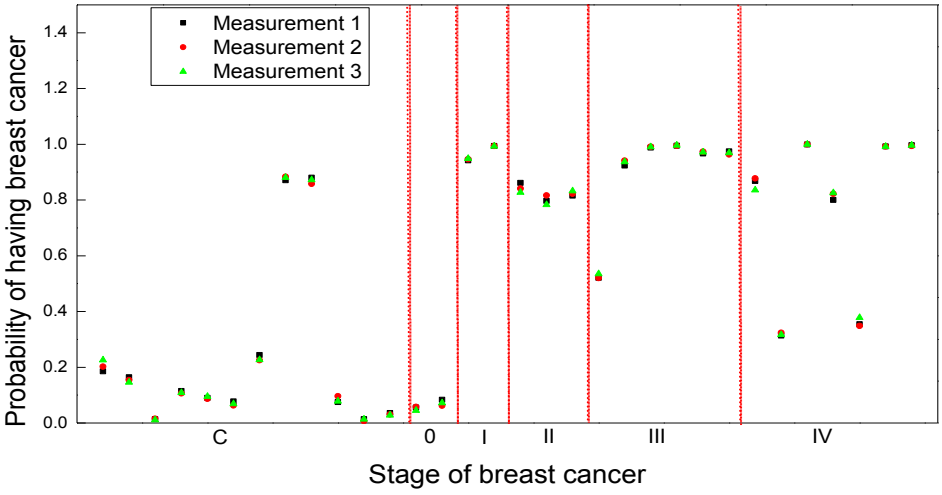


Figure 4-4 Probability of having breast cancer for each patient predicted by logistic regression using individual enzyme pattern as the predictor



4.3 Analysis based on logistic regression with binary response

4.3.1 Use of average MMP as the predictor

A logistic regression model is used to predict the probability of having breast cancer for each person. Patients with breast cancer of staging I, II, III, and IV are grouped together as having breast cancer, whereas the persons without breast cancer or with breast cancer of stage 0 are grouped together as having no breast cancer. When average MMP is used as the predictor, the estimates of parameters are shown in Table 4-1. The probability of having breast cancer for each patient is shown in Table 4-2.

Table 4-1 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-6.5605	2.2377	8.5959	0.0034
MMP	1	11.5520	3.6899	9.8014	0.0017

Table 4-2 Probability of having breast cancer for each patient

Obs	Patient	Age	Staging	MMP	Cancer	Prob
1	B1	41	0	0.36336	No	0.086
2	B2	64	IV	0.37156	Yes	0.094
3	B3	45	0	0.35821	No	0.081
4	B4	59	I	0.75493	Yes	0.897
5	B5	65	III	0.76096	Yes	0.903
6	B6	77	IV	0.77706	Yes	0.918
7	B6a	81	I	0.76101	Yes	0.903
8	B7	78	IV	0.8197	Yes	0.948
9	B8	75	III	0.66058	Yes	0.745
10	B10	59	II	0.64032	Yes	0.698
11	B11	46	III	0.60873	Yes	0.616
12	B12	62	IV	0.62555	Yes	0.661
13	B13	40	IV	0.58121	Yes	0.538
14	B14	40	II	0.79096	Yes	0.929
15	B15	36	III	0.95323	Yes	0.988
16	B16	53	IV	1.2534	Yes	1.000
17	B17	52	III	0.98838	Yes	0.992
18	B18	43	IV	0.77332	Yes	0.915
19	B19	51	II	0.70182	Yes	0.824
20	B20	80	III	0.80806	Yes	0.941
21	C1	47	C	0.45603	No	0.215
22	C2	51	C	0.32572	No	0.057
23	C3	28	C	0.51654	No	0.356
24	C4	32	C	0.61482	No	0.632
25	C5	62	C	0.64015	No	0.697
26	C6	48	C	0.76616	No	0.908
27	C7	49	C	0.33458	No	0.063
28	C8	26	C	0.34843	No	0.073
29	C9	42	C	0.30096	No	0.044
30	C10	60	C	0.29954	No	0.043
31	C11	26	C	0.36683	No	0.089
32	C12	44	C	0.41336	No	0.144

The optimal cut-off probability to predict if the person has breast cancer is found by a ROC curve. Table 4-3 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 4-5 is the ROC curve. The optimal cut-off probability is the one

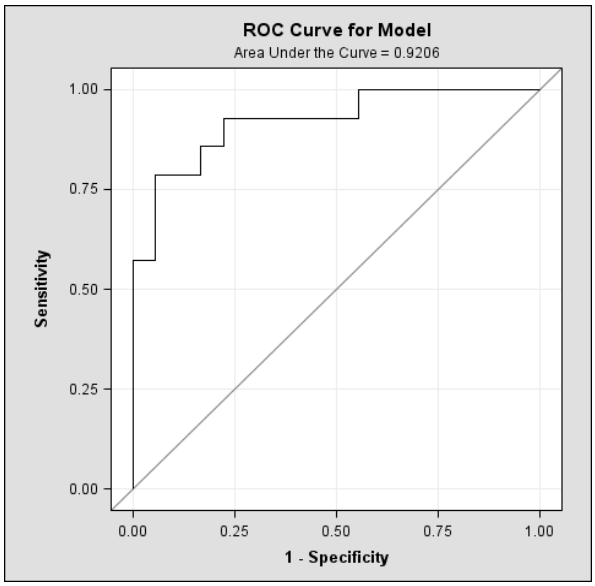
which has the smallest distance to the point (0,1) and at the same time has the largest vertical distance to the line of equality. Based on these criteria, the probability of 0.538 is the optimal cut-off probability to predict if the person has breast cancer. If the predicted probability of a person is above 0.538, this person is predicted to have breast cancer. If the predicted probability of a person is below 0.538, this person is predicted to have no breast cancer. The area under the ROC curve is 0.9206, indicating this test method is excellent.

Table 4-3 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	1.000	0.06	0.00	0.94	0.06
2	0.992	0.11	0.00	0.89	0.11
3	0.988	0.17	0.00	0.83	0.17
4	0.948	0.22	0.00	0.78	0.22
5	0.941	0.28	0.00	0.72	0.28
6	0.929	0.33	0.00	0.67	0.33
7	0.918	0.39	0.00	0.61	0.39
8	0.915	0.44	0.00	0.56	0.44
9	0.908	0.44	0.07	0.56	0.37
10	0.903	0.50	0.07	0.51	0.43
11	0.903	0.56	0.07	0.45	0.49
12	0.897	0.61	0.07	0.40	0.54
13	0.824	0.67	0.07	0.34	0.6
14	0.745	0.72	0.07	0.29	0.65
15	0.698	0.78	0.07	0.23	0.71
16	0.697	0.78	0.14	0.26	0.64
17	0.661	0.83	0.14	0.22	0.69
18	0.632	0.83	0.21	0.27	0.62
19	0.616	0.89	0.21	0.24	0.68
20	0.538	0.94	0.21	0.22	0.73
21	0.356	0.94	0.29	0.29	0.65
22	0.215	0.94	0.36	0.36	0.58
23	0.144	0.94	0.43	0.43	0.51
24	0.094	1.00	0.43	0.43	0.57
25	0.089	1.00	0.50	0.50	0.5
26	0.086	1.00	0.57	0.57	0.43
27	0.081	1.00	0.64	0.64	0.36
28	0.073	1.00	0.71	0.71	0.29
29	0.063	1.00	0.79	0.79	0.21
30	0.057	1.00	0.86	0.86	0.14

31	0.044	1.00	0.93	0.93	0.07
32	0.043	1.00	1.00	1.00	0

Figure 4-5 ROC curve of the model

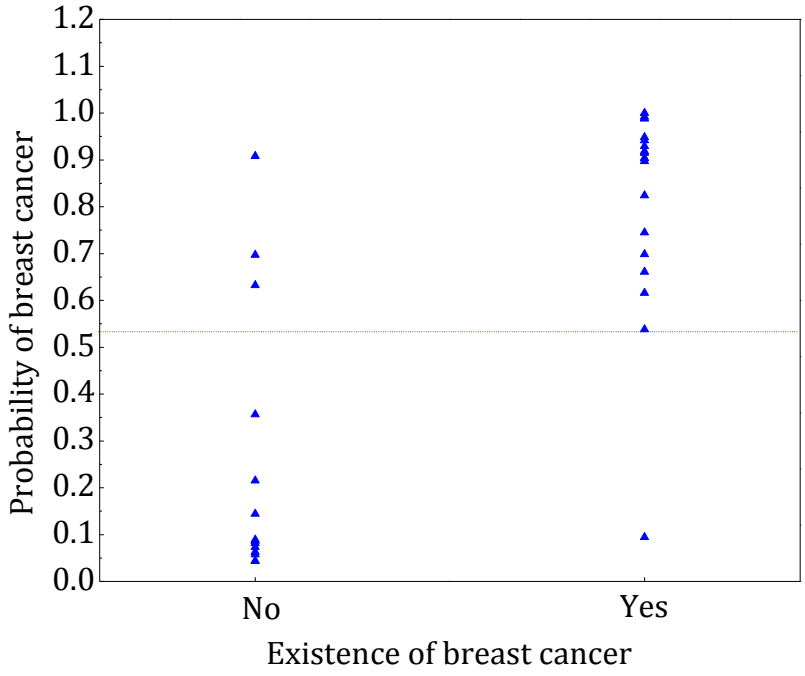


When the cut-off probability is set to be 0.538, the relationship between the predicted existence of breast cancer and the actual diagnosis is shown in Figure 4-6. Three persons in the control group are predicted to have cancer, whereas, one person in the case group is predicted to have no breast cancer. The sensitivity and specificity for this test is 0.94 and 0.79, respectively. Table 4-4 summarizes how many persons are predicted to have breast cancer and how many persons are predicted to have no breast cancer for patients at different stages of breast cancer.

Table 4-4 Prediction of existence of breast cancer for patients at different stages

	Predicted	
	No	Yes
C	9	3
0	2	0
I	0	2
II	0	3
III	0	6
IV	1	6

Figure 4-6 Prediction of existence of cancer based on the optimal cut-off probability



4.3.2 Use of average MMP and age as the predictors

In this section, average MMP and age are used as the predictors to predict if the person has breast cancer. Estimates of parameters are shown in Table 4-5. The probability of having breast cancer for each patient is shown in Table 4-6.

Table 4-5 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Error	Standard Chi-Square	Wald Pr > ChiSq
Intercept	1	-10.5871	3.7763	7.8599	0.0051
MMP	1	10.2277	3.6114	8.0207	0.0046
Age	1	0.0970	0.0565	2.9495	0.0859

Table 4-6 Probability of having breast cancer for each patient

Obs	Patient	Age	Staging	MMP	Cancer	Prob
1	B1	41	0	0.36336	No	0.052
2	B2	64	IV	0.37156	Yes	0.359
3	B3	45	0	0.35821	No	0.072
4	B4	59	I	0.75493	Yes	0.946
5	B5	65	III	0.76096	Yes	0.971
6	B6	77	IV	0.77706	Yes	0.992
7	B6a	81	I	0.76101	Yes	0.994
8	B7	78	IV	0.8197	Yes	0.995

9	B8	75	III	0.66058	Yes	0.969
10	B10	59	II	0.64032	Yes	0.844
11	B11	46	III	0.60873	Yes	0.525
12	B12	62	IV	0.62555	Yes	0.861
13	B13	40	IV	0.58121	Yes	0.318
14	B14	40	II	0.79096	Yes	0.799
15	B15	36	III	0.95323	Yes	0.934
16	B16	53	IV	1.2534	Yes	0.999
17	B17	52	III	0.98838	Yes	0.990
18	B18	43	IV	0.77332	Yes	0.817
19	B19	51	II	0.70182	Yes	0.823
20	B20	80	III	0.80806	Yes	0.996
21	C1	47	C	0.45603	No	0.204
22	C2	51	C	0.32572	No	0.090
23	C3	28	C	0.51654	No	0.070
24	C4	32	C	0.61482	No	0.232
25	C5	62	C	0.64015	No	0.878
26	C6	48	C	0.76616	No	0.870
27	C7	49	C	0.33458	No	0.082
28	C8	26	C	0.34843	No	0.011
29	C9	42	C	0.30096	No	0.031
30	C10	60	C	0.29954	No	0.154
31	C11	26	C	0.36683	No	0.013
32	C12	44	C	0.41336	No	0.110

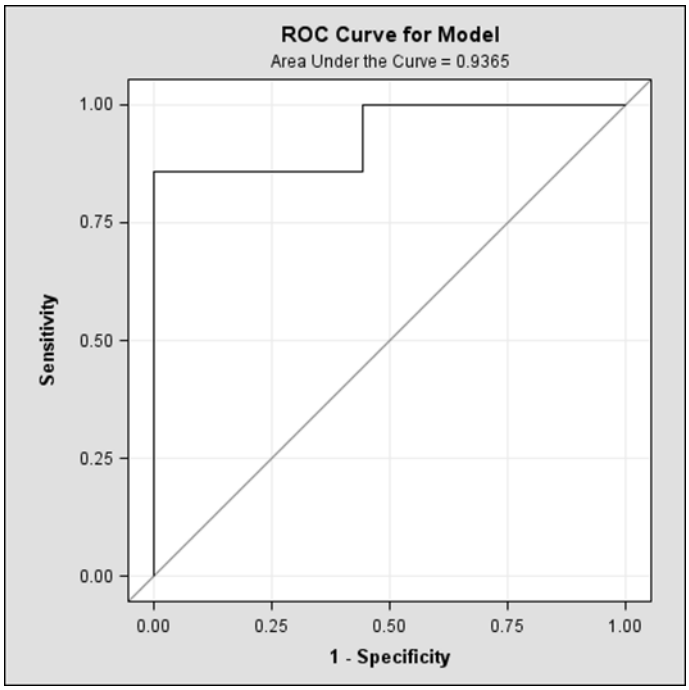
Table 4-7 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 4-7 is the ROC curve. From Table 4-7, it is found that 0.318 is the optimal cut-off probability to predict if the person has breast cancer. If the predicted probability of a person is above 0318, this person is predicted to have breast cancer. If the predicted probability of a person is below 0.3188, this person is predicted to have no breast cancer. The area under the ROC curve is 0.9365, indicating this test method is excellent.

Table 4-7 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	0.999	0.06	0.00	0.94	0.06
2	0.996	0.11	0.00	0.89	0.11
3	0.995	0.17	0.00	0.83	0.17
4	0.994	0.22	0.00	0.78	0.22
5	0.992	0.28	0.00	0.72	0.28
6	0.990	0.33	0.00	0.67	0.33
7	0.971	0.39	0.00	0.61	0.39

8	0.969	0.44	0.00	0.56	0.44
9	0.946	0.50	0.00	0.50	0.5
10	0.934	0.56	0.00	0.44	0.56
11	0.878	0.56	0.07	0.45	0.49
12	0.870	0.56	0.14	0.47	0.42
13	0.861	0.61	0.14	0.41	0.47
14	0.843	0.67	0.14	0.36	0.53
15	0.823	0.72	0.14	0.31	0.58
16	0.816	0.78	0.14	0.26	0.64
17	0.799	0.83	0.14	0.22	0.69
18	0.525	0.89	0.14	0.18	0.75
19	0.359	0.94	0.14	0.15	0.8
20	0.318	1.00	0.14	0.14	0.86
21	0.232	1.00	0.21	0.21	0.79
22	0.203	1.00	0.29	0.29	0.71
23	0.154	1.00	0.36	0.36	0.64
24	0.110	1.00	0.43	0.43	0.57
25	0.090	1.00	0.50	0.50	0.5
26	0.082	1.00	0.57	0.57	0.43
27	0.072	1.00	0.64	0.64	0.36
28	0.070	1.00	0.71	0.71	0.29
29	0.052	1.00	0.79	0.79	0.21
30	0.031	1.00	0.86	0.86	0.14
31	0.013	1.00	0.93	0.93	0.07
32	0.011	1.00	1.00	1.00	0

Figure 4-7 ROC curve of the model



When the cut-off probability is set to be 0.318, the relationship between the predicted existence of breast cancer and the actual diagnosis is shown in Figure 4-8. One person in the control group is predicted to have breast cancer, whereas, no person in the case group is predicted to have no breast cancer. The sensitivity and specificity for this test is 1.00 and 0.86, respectively. Table 4-8 summarizes how many persons are predicted to have breast cancer and how many persons are predicted to have no breast cancer for patients at different stages.

Figure 4-8 Prediction of existence of cancer based on the optimal cut-off probability

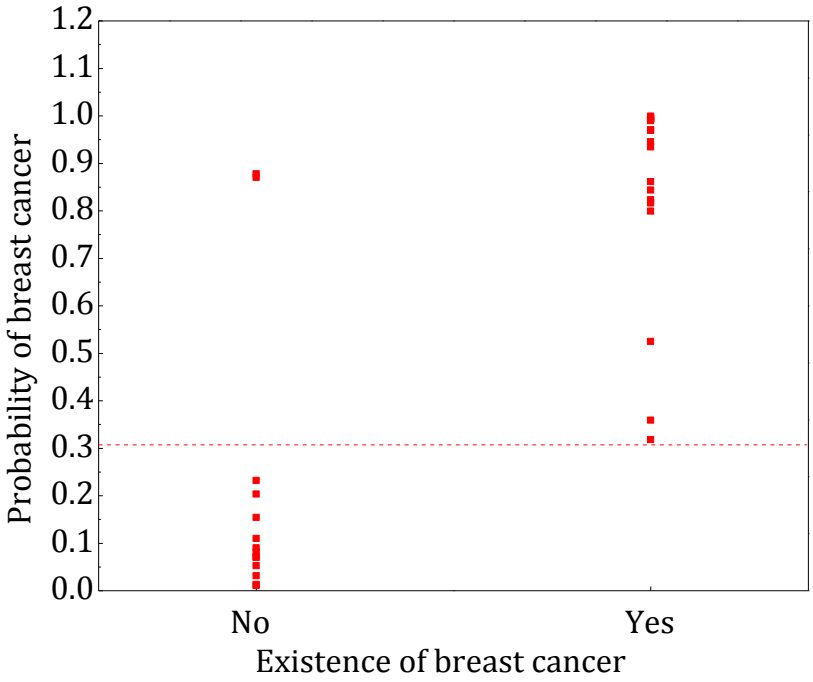


Table 4-8 Prediction of existence of breast cancer for patients at different stages

	Predicted	
	No	Yes
C	10	2
0	2	0
I	0	2
II	0	3
III	0	6
IV	0	7

4.3.3 Comparison of models in section 4.2.1 and 4.2.2

Table 4-9 compares the probabilities calculated by the models in section 4.2.1 (average MMP is used as the predictor) and section 4.2.2 (average MMP and age are used as the predictors). The difference between the probabilities is relatively large when the age of the patient is far from the average age of the patients involved in the test. When the age of the patient is close to the average age, the difference is relatively small, indicating age plays a role in the analysis.

Table 4-9 Comparison of probabilities of having breast cancer for each patient between the two models in section 4.2.1 and 4.2.2

Obs	Patient	Age	Staging	MMP	Cancer	Prob1	Prob2
1	B1	41	0	0.36336	No	0.086	0.052
2	B2	64	IV	0.37156	Yes	0.094	0.359
3	B3	45	0	0.35821	No	0.081	0.072
4	B4	59	I	0.75493	Yes	0.897	0.946
5	B5	65	III	0.76096	Yes	0.903	0.971
6	B6	77	IV	0.77706	Yes	0.918	0.992
7	B6a	81	I	0.76101	Yes	0.903	0.994
8	B7	78	IV	0.8197	Yes	0.948	0.995
9	B8	75	III	0.66058	Yes	0.745	0.969
10	B10	59	II	0.64032	Yes	0.698	0.844
11	B11	46	III	0.60873	Yes	0.616	0.525
12	B12	62	IV	0.62555	Yes	0.661	0.861
13	B13	40	IV	0.58121	Yes	0.538	0.318
14	B14	40	II	0.79096	Yes	0.929	0.799
15	B15	36	III	0.95323	Yes	0.988	0.934
16	B16	53	IV	1.2534	Yes	1	0.999
17	B17	52	III	0.98838	Yes	0.992	0.99
18	B18	43	IV	0.77332	Yes	0.915	0.817
19	B19	51	II	0.70182	Yes	0.824	0.823
20	B20	80	III	0.80806	Yes	0.941	0.996
21	C1	47	C	0.45603	No	0.215	0.204
22	C2	51	C	0.32572	No	0.057	0.09
23	C3	28	C	0.51654	No	0.356	0.07
24	C4	32	C	0.61482	No	0.632	0.232
25	C5	62	C	0.64015	No	0.697	0.878
26	C6	48	C	0.76616	No	0.908	0.87
27	C7	49	C	0.33458	No	0.063	0.082
28	C8	26	C	0.34843	No	0.073	0.011
29	C9	42	C	0.30096	No	0.044	0.031
30	C10	60	C	0.29954	No	0.043	0.154

31	C11	26	C	0.36683	No	0.089	0.013
32	C12	44	C	0.41336	No	0.144	0.11

Table 4-10 summarizes the differences of the two models in the diagnostic test of breast cancer. Including age as the predictor increases the area under the ROC curve, decreases the cut-off probability, and increases sensitivity and specificity of the test. The p-value for the predictor, average MMP, is decreased. The p-value for the predictor age is 0.0859, indicating that age has a marginal influence on the prediction. Comparison of predicted accuracy between the two models is shown in Figure 4-9.

Figure 4-9 Prediction of existence of cancer based on the optimal cut-off probability for the models used in section 4.2.1 and 4.2.2

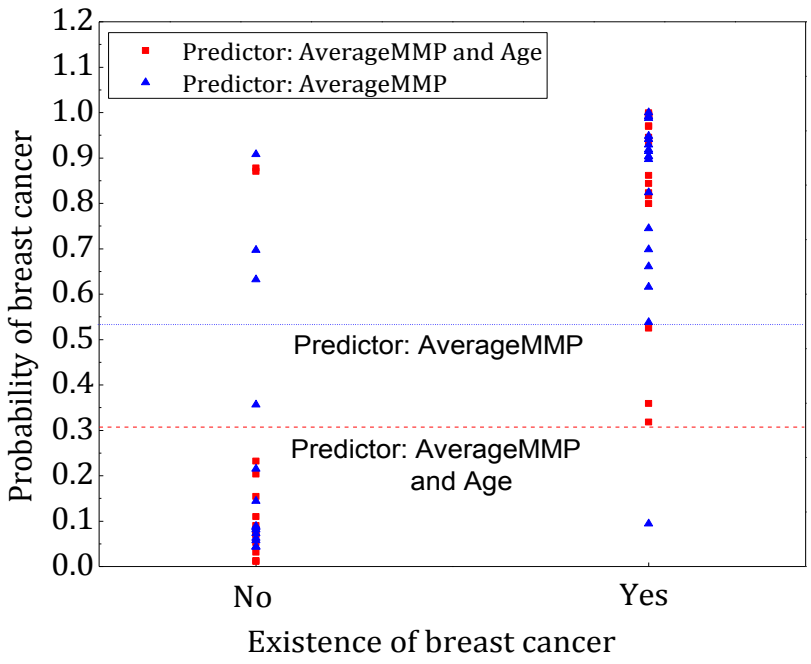


Table 4-10 Comparison of models in section 4.2.1 and 4.2.2

Predictor	Area under the ROC curve	Threshold	Sensi	Speci	P-value	
					MMP	Age
Average MMP	0.9206	0.538	0.94	0.79	0.0017	N/A
Average MMP+Age	0.9365	0.318	1	0.86	0.0046	0.0859

4.4 Analysis based on multicategory logistic model

In this section, average MMP and age are used as predictors to predict the probabilities of each stage of cancer for different persons. Analysis of effects of different parameters is shown in Table 4-11. Estimates of parameters are shown in Table 4-12. Table 4-12 lists the intercepts and coefficients of the four equations fit for the four cancer stages, I, II, III, and IV. The intercept and coefficient of the equation for the control group (including the healthy persons and the patients with breast cancer of stage 0) are set to be 0 by default. Probabilities of each stage of breast cancer for different persons are calculated and shown in Table 4-13. For example, the first line of Table 4-13 indicates the first patient has a breast cancer at stage 0. Based on the diagnostic test, the probability that he or she is at stage 0 or has no cancer is 0.946. The probability that he or she is in stage I, II, III, or IV is 0.001, 0.026, 0.009, and 0.019, respectively. The p-values of average MMP and age are 0.0858 and 0.3389, indicating average MMP has a marginal effect on stage of breast cancer, whereas age has little effect on stage of breast cancer.

Table 4-11 Analysis of Effects

Wald Effect	DF	Chi-Square	Pr> ChiSq
average MMP	4	8.1626	0.0858
Age	4	4.531	0.3389

Table 4-12 Intercepts and coefficients of the logistic regression equations fit for the four cancer stages

Standard Parameter	Wald Staging	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	IV	1	-12.1065	4.4202	7.5017	0.0062
Intercept	III	1	-13.4324	4.7881	7.8701	0.005
Intercept	II	1	-9.1344	4.3973	4.3151	0.0378
Intercept	I	1	-17.8201	7.3013	5.957	0.0147
average MMP	IV	1	10.1164	4.0257	6.3148	0.012
average MMP	III	1	11.6952	4.3402	7.2609	0.007
average MMP	II	1	8.9862	4.3099	4.3472	0.0371
average MMP	I	1	10.8368	6.4113	2.857	0.091
Age	IV	1	0.1101	0.0632	3.0387	0.0813
Age	III	1	0.1094	0.0653	2.807	0.0939
Age	II	1	0.0552	0.0672	0.6757	0.4111
Age	I	1	0.1704	0.0875	3.7937	0.0514

Table 4-13 Probabilities of breast cancer in each stage for each patient

Obs	Age	Staging	average MMP	Prob at stage I	Prob at stage II	Prob at stage III	Prob at stage IV	Prob at stage 0 or no cnacer
1	41	0	0.36336	0.001	0.026	0.009	0.019	0.946
2	64	IV	0.37156	0.036	0.067	0.080	0.175	0.643
3	45	0	0.35821	0.002	0.030	0.012	0.027	0.929
4	59	I	0.75493	0.080	0.131	0.336	0.401	0.053
5	65	III	0.76096	0.118	0.096	0.347	0.412	0.026
6	77	IV	0.77706	0.232	0.046	0.331	0.386	0.006
7	81	I	0.76101	0.280	0.036	0.309	0.371	0.004
8	78	IV	0.8197	0.244	0.040	0.341	0.372	0.003
9	75	III	0.66058	0.203	0.062	0.296	0.414	0.024
10	59	II	0.64032	0.069	0.139	0.262	0.374	0.157
11	46	III	0.60873	0.017	0.158	0.136	0.202	0.488
12	62	IV	0.62555	0.085	0.125	0.266	0.389	0.136
13	40	IV	0.58121	0.006	0.125	0.072	0.111	0.686
14	40	II	0.79096	0.018	0.247	0.250	0.278	0.206
15	36	III	0.95323	0.017	0.274	0.346	0.297	0.066
16	53	IV	1.2534	0.060	0.078	0.559	0.302	0.001
17	52	III	0.98838	0.056	0.133	0.440	0.361	0.010
18	43	IV	0.77332	0.023	0.229	0.260	0.298	0.190
19	51	II	0.70182	0.040	0.181	0.261	0.336	0.183
20	80	III	0.80806	0.268	0.035	0.328	0.365	0.003
21	47	C	0.45603	0.006	0.070	0.042	0.079	0.803
22	51	C	0.32572	0.003	0.031	0.016	0.037	0.913
23	28	C	0.51654	0.001	0.048	0.012	0.021	0.919
24	32	C	0.61482	0.003	0.120	0.049	0.071	0.757
25	62	C	0.64015	0.087	0.124	0.275	0.394	0.119
26	48	C	0.76616	0.035	0.200	0.292	0.339	0.134
27	49	C	0.33458	0.003	0.030	0.014	0.033	0.920
28	26	C	0.34843	0.000	0.010	0.001	0.003	0.985
29	42	C	0.30096	0.001	0.016	0.005	0.011	0.967
30	60	C	0.29954	0.011	0.037	0.029	0.072	0.851
31	26	C	0.36683	0.000	0.012	0.002	0.004	0.982
32	44	C	0.41336	0.003	0.045	0.020	0.041	0.891

For each person, the stage of cancer is predicted to be the one with the largest probability. As a summary, table 4-14 lists the numbers of persons predicted to be breast cancers of different stages.

Table 4-14 Prediction of staging of breast cancer for patients in different staging's of breast cancer

Staging	C	III	IV	Sum
C	10		2	12
0	2			2
I			2	2
II			3	3
III	1	3	2	6
IV	2	2	3	7
Sum	15	5	12	32

4.5 Analysis based on CART

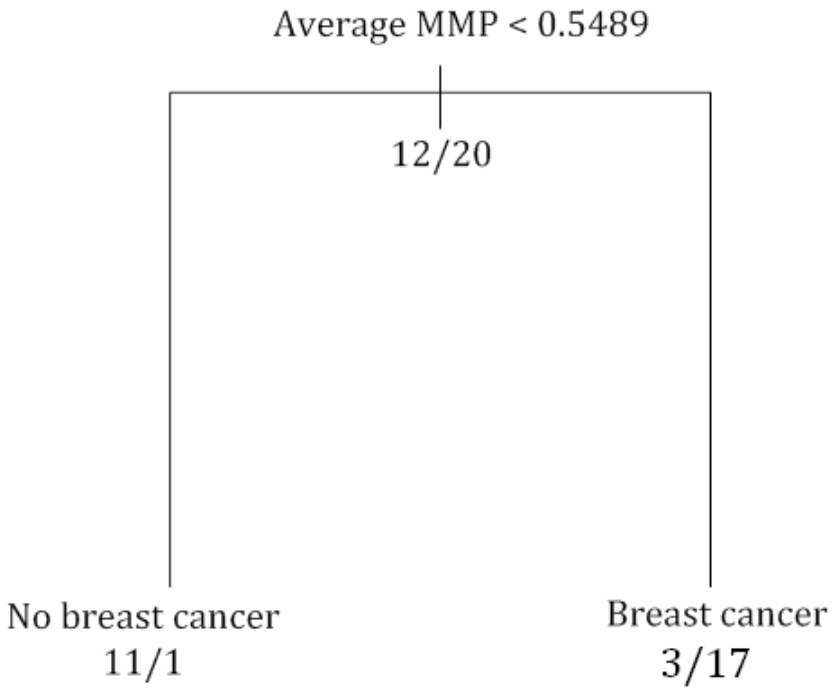
The data are also analyzed by classification tree. These patients are divided into two groups based on the method used in section 4.3. If the patient has no breast cancer or has breast cancer at stage 0, he is in the no breast cancer group. If the patient has breast cancer at stage I, II, III or IV, he is in the breast cancer group. The output is shown in Figure 4-10. 12 persons have average MMP below 0.5489 and are predicted to have no breast cancer. Out of them, 1 person has breast cancer. 20 persons are predicted to have breast cancer. Out of them, 3 persons have no breast cancer. If the average MMP is greater than 0.5489, the patient is predicted to have breast cancer. On the contrary, if the average MMP of the individual is smaller than 0.5489, the individual is predicted to have no breast cancer. The prediction for each patient is in Table 4-16.

Table 4-15 Predicted breast cancer condition by tree model

Patient	Age	Staging	MMP11	MMP12	MMP13	Average MMP	Cancer
B1	41	0	0.352169	0.377403	0.360494	0.3633552	No
B2	64	IV	0.353283	0.377405	0.384004	0.37156407	No
B3	45	0	0.357591	0.34873	0.3683	0.35820721	No
B4	59	I	0.741549	0.75993	0.763322	0.75493378	Yes
B5	65	III	0.743864	0.774038	0.764992	0.76096495	Yes
B6	77	IV	0.774821	0.782392	0.773965	0.77705943	Yes
B6a	81	I	0.760201	0.773043	0.749784	0.7610094	Yes
B7	78	IV	0.827511	0.80484	0.826743	0.81969806	Yes
B8	75	III	0.668552	0.653095	0.660093	0.66057994	Yes
B10	59	II	0.644745	0.643923	0.632294	0.6403208	Yes
B11	46	III	0.59722	0.608893	0.620085	0.6087327	Yes

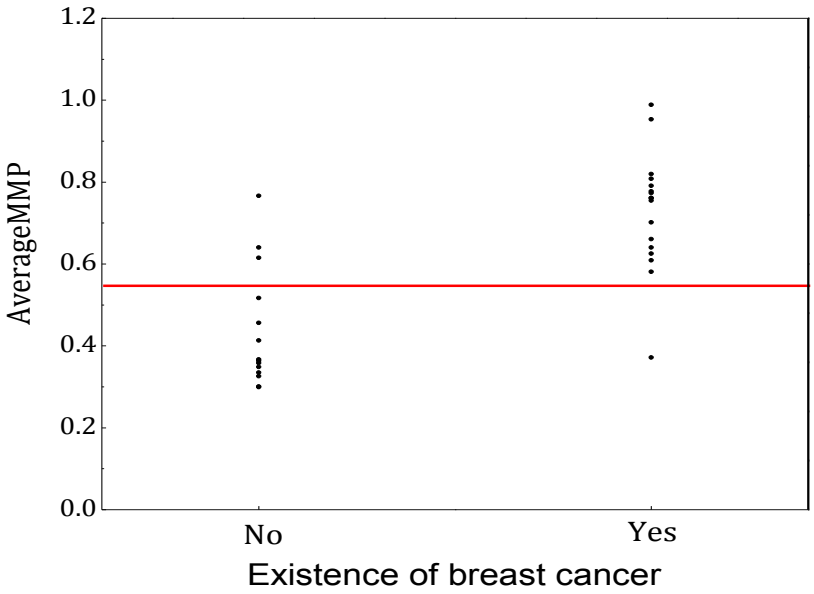
B12	62	IV	0.62226	0.645563	0.608827	0.62555011	Yes
B13	40	IV	0.56818	0.584877	0.590558	0.58120536	Yes
B14	40	II	0.784446	0.799234	0.789203	0.79096106	Yes
B15	36	III	0.935575	0.958493	0.965632	0.95323343	Yes
B16	53	IV	1.212854	1.269944	1.277404	1.25340073	Yes
B17	52	III	0.972538	1.003827	0.988784	0.98838307	Yes
B18	43	IV	0.757089	0.776089	0.786774	0.77331739	Yes
B19	51	II	0.688673	0.703384	0.713394	0.7018173	Yes
B20	80	III	0.796674	0.809744	0.817749	0.80805586	Yes
C1	47	C	0.43033	0.459904	0.477849	0.456028	No
C2	51	C	0.308407	0.328867	0.339895	0.32572294	No
C3	28	C	0.514881	0.504738	0.529989	0.51653638	No
C4	32	C	0.611661	0.608947	0.623849	0.61481918	Yes
C5	62	C	0.624123	0.650948	0.645373	0.64014824	Yes
C6	48	C	0.768959	0.756474	0.773044	0.76615914	Yes
C7	49	C	0.307784	0.357486	0.338473	0.33458102	No
C8	26	C	0.355648	0.300899	0.388749	0.34843235	No
C9	42	C	0.297403	0.30499	0.300494	0.30096246	No
C10	60	C	0.288947	0.309804	0.299877	0.29954283	No
C11	26	C	0.355894	0.377849	0.36674	0.3668279	No
C12	44	C	0.403832	0.413395	0.422849	0.41335884	No

Figure 4-10 Prediction of breast cancer condition based on CART model



The prediction for patients with breast cancer and without breast cancer is illustrated in Figure 4-11. It can be seen that one person with breast cancer is predicted to have no breast cancer, whereas, three patients without breast cancer are predicted to have breast cancer.

Figure 4-11 Scatter plot of enzyme pattern vs existence of breast cancer



It needs to be noted that the number of wrongly predicted patients is the same as the number of wrongly predicted patients obtained by the logistic regression model when average MMP is used as the predictor and this grouping is used (model in section 3.4.1).

Chapter 5 - Assessing Diagnostic Test for Lung Cancer (data MMP1)

5.1 Overview of the data

The nanoparticles used in Chapter 4 are also used to detect lung cancer by examining the enzyme pattern in the blood of a patient. Information on enzyme activity, patients' ages and stages of cancer were recorded. Relationships between average MMP, stage of cancer, and age of patient are shown in Figure 5-1, Figure 5-2, and Figure 5-3. From these boxplots, it can be seen that the patients with lung cancer of stages I, II, and III have higher average MMP than the persons in the control group. But the differences among average MMP of patients with lung cancer of stage I, II, and III is not remarkable. Age of patient does not show a significant relationship with average MMP or stage of lung cancer.

Figure 5-1 Boxplot of average MMP vs staging of cancer

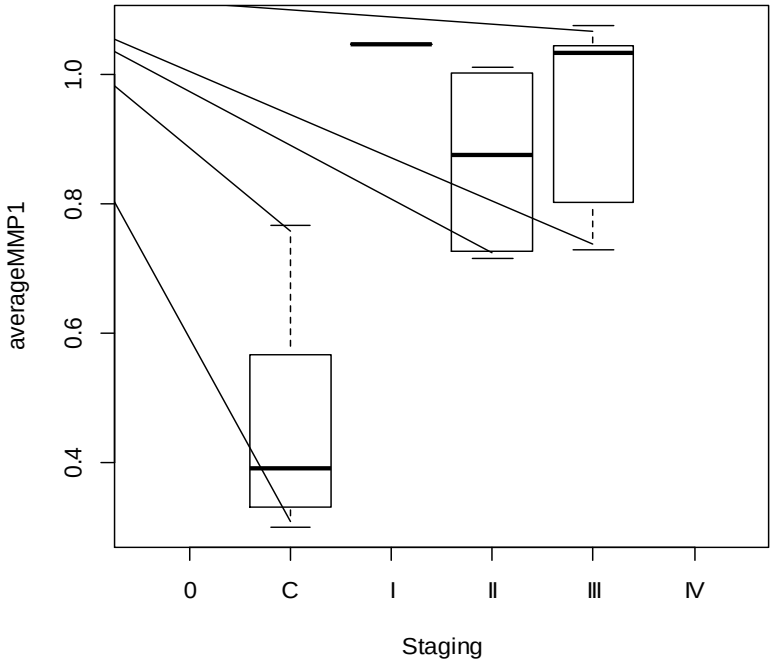


Figure 5-2 Scatterplot of average MMP vs age of patient

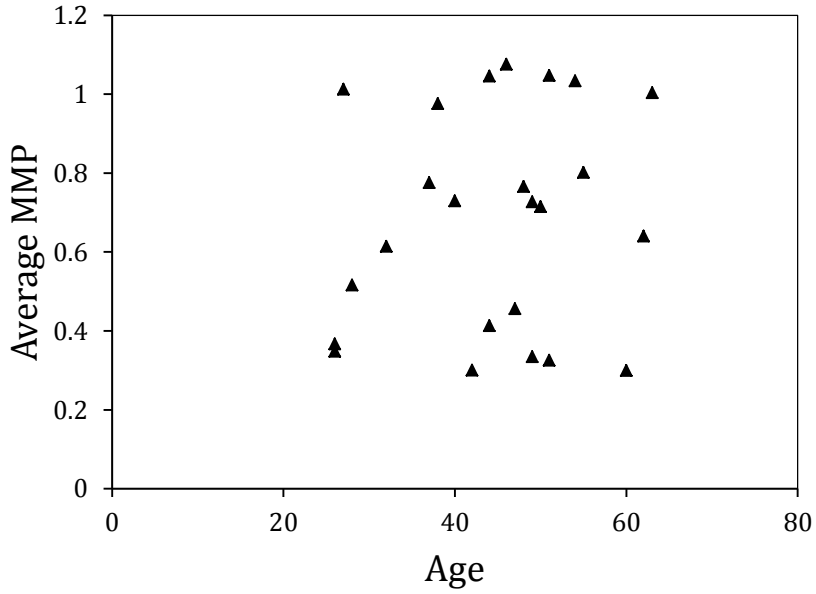
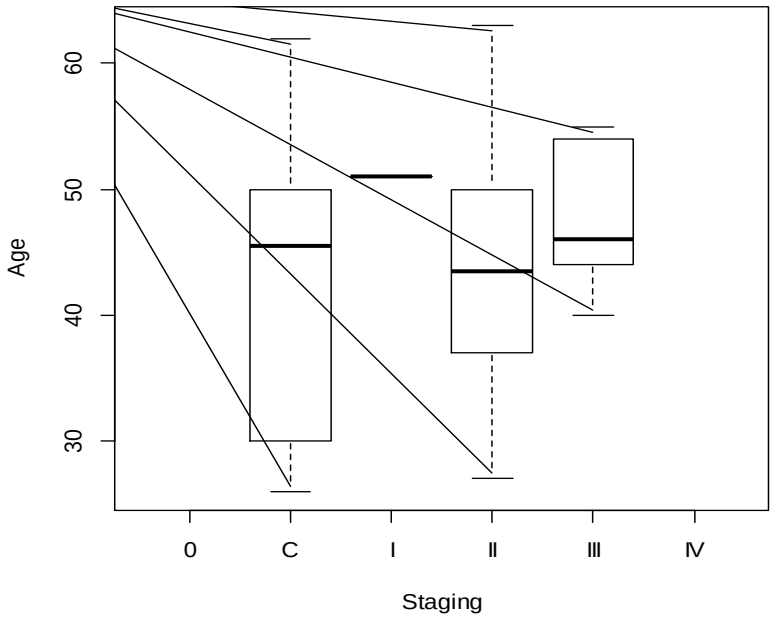


Figure 5-3 Boxplot of age of patient vs staging of cancer



5.2 Comparison of three measurements

The enzyme pattern of each person was tested three times on the different days. Figure 5-4 shows the results of the three measurements. It can be seen that the experimental error is very small. In addition, a logistic regression model is also used to predict the probability of

having lung cancer for each person by using the individual enzyme pattern instead of the average enzyme pattern of the three measurements. If the person has lung cancer of stage I, II and III, this person belongs to the "having lung cancer" group. If the person has no lung cancer, the person belongs to the "having no lung cancer" group. The probability of having lung cancer for each patient is shown in Figure 5-5, which shows that for most the persons, there is no noticeable difference between the three probabilities, but three indicate some wide variability in Figure 5.5. As a result, in this chapter, the models are built based on the average value of these three measurements.

Figure 5-4 Scatter plot of the three enzyme patterns of different patients

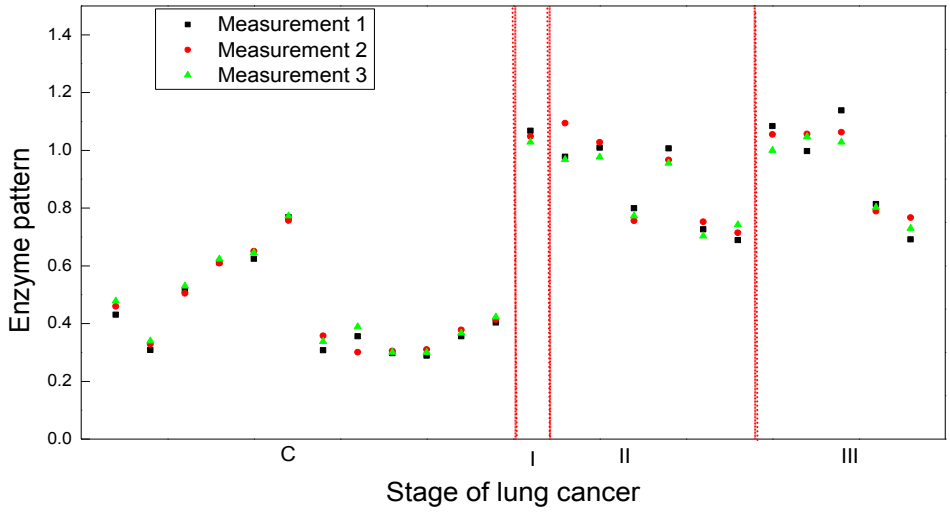
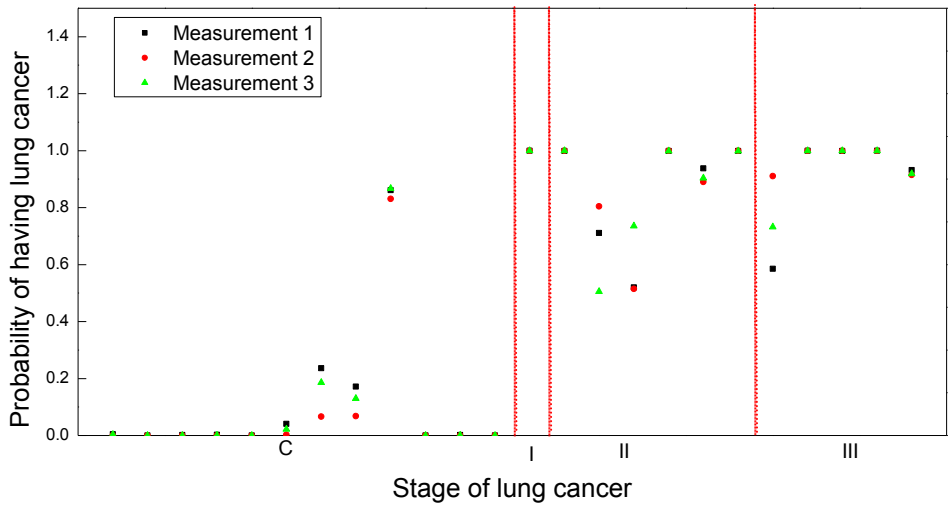


Figure 5-5 Probability of having breast cancer for each patient predicted by logistic regression using individual enzyme pattern as the predictor



5.3 Analysis based on logistic regression with binary response

5.3.1 Use of average MMP as the predictor

A logistic regression model is used to predict the probability of having lung cancer for each person. Patients with lung cancer of staging I, II, and III are grouped together as having lung cancer, whereas the persons without lung cancer are grouped together as having no lung cancer. When average MMP is used as the predictor, the estimates of parameters are shown in Table 5-1. Probability of having lung cancer for each patient is shown in Table 5-2.

Table 5-1 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-18.9824	12.3881	2.3480	0.1254
average MMP	1	27.2166	17.2605	2.4864	0.1148

Table 5-2 Probability of having lung cancer for each patient

Obs	Patient	Staging	Average MMP	Cancer	Prob
1	C1	C	0.45603	No	0.001
2	C2	C	0.32572	No	0.000
3	C3	C	0.51654	No	0.007
4	C4	C	0.61482	No	0.095
5	C5	C	0.64015	No	0.174
6	C6	C	0.76616	No	0.866
7	C7	C	0.33458	No	0.000
8	C8	C	0.34843	No	0.000
9	C9	C	0.30096	No	0.000
10	C10	C	0.29954	No	0.000
11	C11	C	0.36683	No	0.000
12	C12	C	0.41336	No	0.000
13	L1	II	1.01316	Yes	1.000
14	L2	III	1.04613	Yes	1.000
15	L3	II	1.00416	Yes	1.000
16	L4	III	1.03366	Yes	1.000
17	L5	II	0.77618	Yes	0.895
18	L6	III	1.07635	Yes	1.000
19	L7	I	1.04785	Yes	1.000
20	L8	II	0.97629	Yes	0.999
21	L9	III	0.80185	Yes	0.945
22	L10	II	0.72741	Yes	0.693
23	L11	III	0.72932	Yes	0.704
24	L12	II	0.71529	Yes	0.619

The optimal cut-off probability to predict if the person has lung cancer is found by a ROC curve. Table 5-3 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 5-6 is the ROC curve. The optimal point is the one which has the smallest distance to the point (0, 1) and at the same time has the largest vertical distance to the line of equality. Based on these criteria, the probability of 0.619 is the optimal cut-off probability to predict if the person has lung cancer or not. If the predicted probability of a person is above 0619, this person is predicted to have lung cancer. If the predicted probability of a person is below 0.619, this person is predicted to have no lung cancer. The area under the ROC curve is 0.9792, indicating this test method is excellent.

Figure 5-6 ROC curve of the model

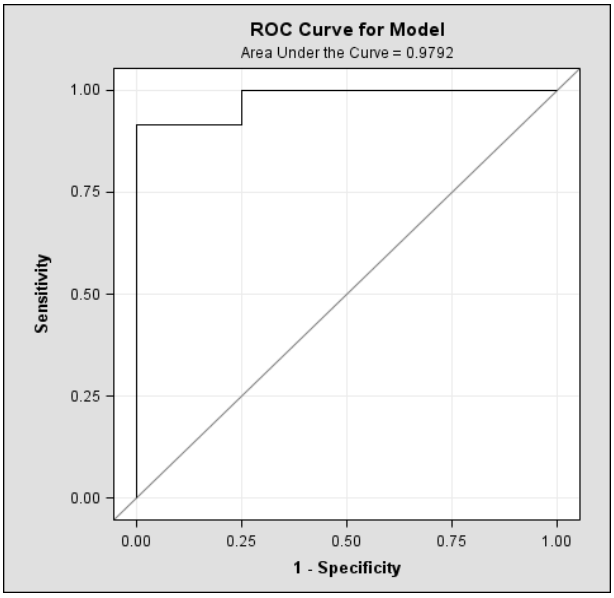


Table 5-3 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	1.000	0.08	0.00	0.92	0.08
2	1.000	0.17	0.00	0.83	0.17
3	1.000	0.25	0.00	0.75	0.25
4	1.000	0.33	0.00	0.67	0.33
5	1.000	0.42	0.00	0.58	0.42
6	1.000	0.50	0.00	0.50	0.5
7	0.999	0.58	0.00	0.42	0.58

8	0.945	0.67	0.00	0.33	0.67
9	0.895	0.75	0.00	0.25	0.75
10	0.866	0.75	0.08	0.26	0.67
11	0.704	0.83	0.08	0.19	0.75
12	0.693	0.92	0.08	0.12	0.84
13	0.619	1.00	0.08	0.08	0.92
14	0.174	1.00	0.17	0.17	0.83
15	0.095	1.00	0.25	0.25	0.75
16	0.007	1.00	0.33	0.33	0.67
17	0.001	1.00	0.42	0.42	0.58
18	0.000	1.00	0.50	0.50	0.5
19	0.000	1.00	0.58	0.58	0.42
20	0.000	1.00	0.67	0.67	0.33
21	0.000	1.00	0.75	0.75	0.25
22	0.000	1.00	0.83	0.83	0.17
23	0.000	1.00	0.92	0.92	0.08
24	0.000	1.00	1.00	1.00	0

When the cut-off probability is set to be 0.619, the relationship between the predicted existence of lung cancer and the actual diagnosis is shown in Figure 5-7. It can be seen that no patient with lung cancer is predicted to have lung cancer, whereas, one person without lung cancer is predicted to have lung cancer. The sensitivity and specificity for this test is 1.00 and 0.92, respectively. Table 5-4 summarizes how many persons are predicted to have lung cancer and how many persons are predicted to have no lung cancer for patients at different stages.

Figure 5-7 Prediction of existence of cancer based on the optimal cut-off probability

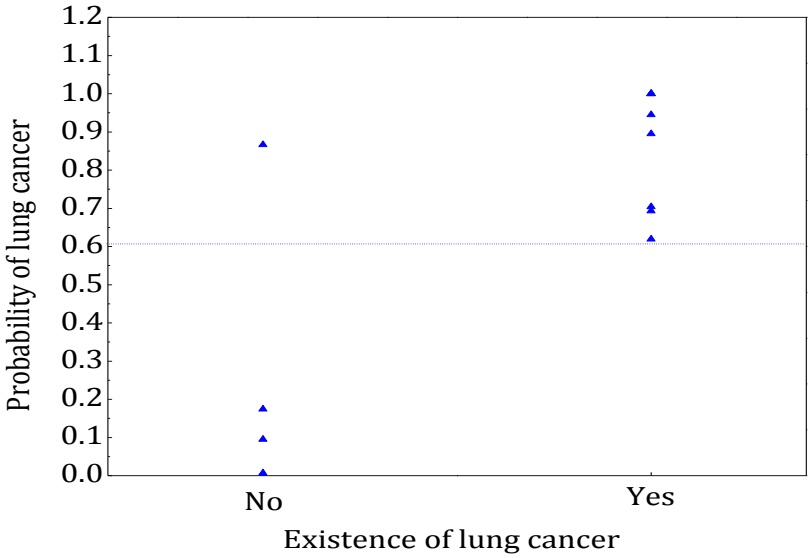


Table 5-4Prediction of existence of lung cancer for patients at different stages

	Predicted	
	No	Yes
C	11	1
I	0	1
II	0	6
III	0	5

5.3.2 Use of average MMP and Age as the predictor

In this section, average MMP and age are used as the predictors to predict if the person has lung cancer. Estimates of parameters are shown in Table 5-5. Probability of having lung cancer for each patient is shown in Table 5-6.

Table 5-5 Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	1	-17.6394	12.0810	2.1319	0.1443
average MMP	1	27.7137	17.0313	2.6478	0.1037
Age	1	-0.0364	0.1180	0.0954	0.7574

Table 5-6 Probability of having lung cancer for each patient

Obs	Patient	Age	Staging	average MMP	Cancer	Prob
1	C1	47	C	0.45603	No	0.001
2	C2	51	C	0.32572	No	0.000
3	C3	28	C	0.51654	No	0.013
4	C4	32	C	0.61482	No	0.146
5	C5	62	C	0.64015	No	0.104
6	C6	48	C	0.76616	No	0.864
7	C7	49	C	0.33458	No	0.000
8	C8	26	C	0.34843	No	0.000
9	C9	42	C	0.30096	No	0.000
10	C10	60	C	0.29954	No	0.000
11	C11	26	C	0.36683	No	0.000
12	C12	44	C	0.41336	No	0.000
13	L1	27	II	1.01316	Yes	1.000
14	L2	44	III	1.04613	Yes	1.000
15	L3	63	II	1.00416	Yes	1.000
16	L4	54	III	1.03366	Yes	1.000
17	L5	37	II	0.77618	Yes	0.926
18	L6	46	III	1.07635	Yes	1.000
19	L7	51	I	1.04785	Yes	1.000
20	L8	38	II	0.97629	Yes	1.000
21	L9	55	III	0.80185	Yes	0.930

22	L10	49	II	0.72741	Yes	0.676
23	L11	40	III	0.72932	Yes	0.753
24	L12	50	II	0.71529	Yes	0.590

Table 5-7 lists sensitivity and 1-specificity calculated under different cut-off probabilities and Figure 5-8 is the ROC curve. From Table 5-7, it is found that 0.589 is the optimal cut-off probability to predict if the person has lung cancer. If the predicted probability of a person is above 0.589, this person is predicted to have lung cancer. If the predicted probability of a person is below 0.589, this person is predicted to have no lung cancer. The area under the ROC curve is 0.9792, indicating this test method is excellent.

Figure 5-8 ROC curve of the model

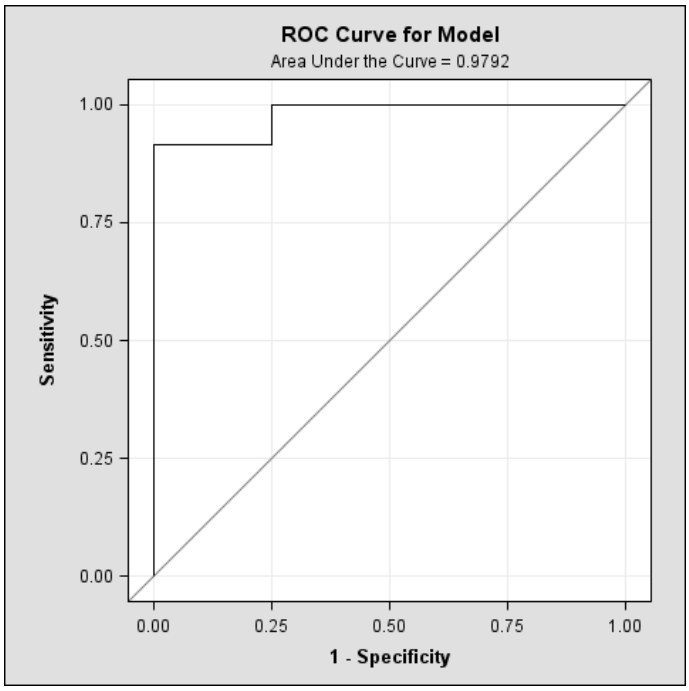


Table 5-7 Relationship between cut-off probability and sensitivity, 1-specificity, distance to the point of (0, 1), and Youden index

Obs	_PROB_	_SENSIT_	_1MSPEC_	DIST to (0,1)	Youden index
1	1.000	0.08	0.00	0.92	0.08
2	1.000	0.17	0.00	0.83	0.17
3	1.000	0.25	0.00	0.75	0.25
4	1.000	0.33	0.00	0.67	0.33
5	1.000	0.42	0.00	0.58	0.42

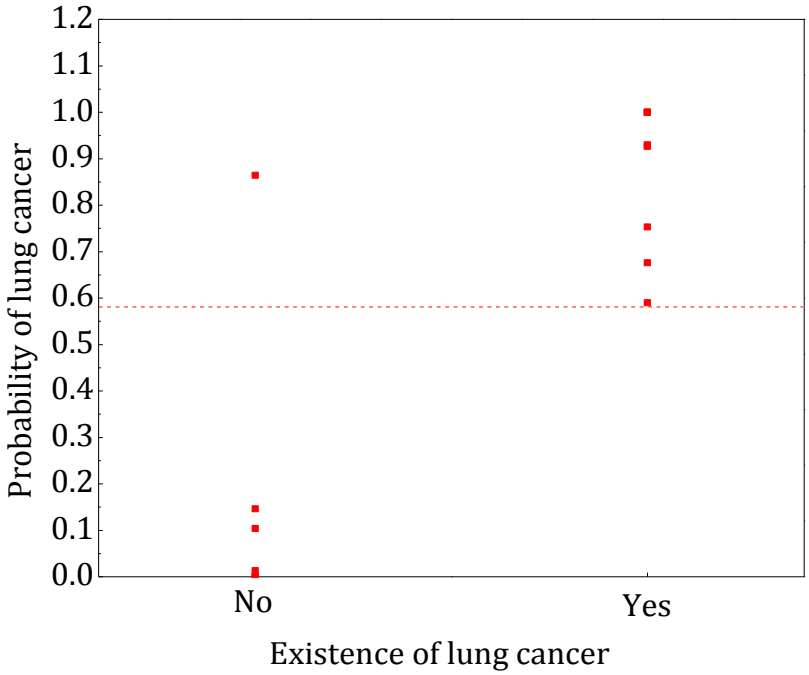
6	1.000	0.50	0.00	0.50	0.50
7	1.000	0.58	0.00	0.42	0.58
8	0.929	0.67	0.00	0.33	0.67
9	0.926	0.75	0.00	0.25	0.75
10	0.863	0.75	0.08	0.26	0.67
11	0.753	0.83	0.08	0.19	0.75
12	0.676	0.92	0.08	0.12	0.83
13	0.589	1.00	0.08	0.08	0.92
14	0.146	1.00	0.17	0.17	0.83
15	0.104	1.00	0.25	0.25	0.75
16	0.013	1.00	0.33	0.33	0.67
17	0.001	1.00	0.42	0.42	0.58
18	0.000	1.00	0.50	0.50	0.50
19	0.000	1.00	0.58	0.58	0.42
20	0.000	1.00	0.67	0.67	0.33
21	0.000	1.00	0.75	0.75	0.25
22	0.000	1.00	0.83	0.83	0.17
23	0.000	1.00	0.92	0.92	0.08
24	0.000	1.00	1.00	1.00	0.00

When the cut-off probability is set to be 0.589, the relationship between the predicted existence of lung cancer and the actual diagnosis is shown in Figure 5-9. One person in the control group is predicted to have lung cancer, whereas, no person in the case group is predicted to have no lung cancer. The sensitivity and specificity for this test is 1.00 and 0.92, respectively. Table 5-8 summarizes how many persons are predicted to have lung cancer and how many persons are predicted to have no lung cancer for patients in different stages.

Table 5-8 Prediction of existence of lung cancer for patients at different stages

	Predicted	
	No	Yes
C	11	1
I	0	1
II	0	6
III	0	5

Figure 5-9 Prediction of existence of cancer based on the optimal cut-off probability



5.3.3 Comparison of models in section 5.2.1 and 5.2.2

Table 5-9 compares the probabilities calculated by the models in section 5.2.1 (average MMP is used as the predictor) and section 5.2.2 (average MMP and age are used as the predictors). The probabilities calculated by the two models are similar for all the persons.

Table 5-9 Comparison of probabilities of having lung cancer for each patient between the two models in section 5.2.1 and 5.2.2

Obs	Patient	Age	Staging	average MMP	Cancer	Prob1	Prob2
1	C1	47	C	0.45603	No	0.001	0.001
2	C10	60	C	0.29954	No	0.000	0.000
3	C11	26	C	0.36683	No	0.000	0.000
4	C12	44	C	0.41336	No	0.000	0.000
5	C2	51	C	0.32572	No	0.000	0.000
6	C3	28	C	0.51654	No	0.007	0.013
7	C4	32	C	0.61482	No	0.095	0.146
8	C5	62	C	0.64015	No	0.174	0.104
9	C6	48	C	0.76616	No	0.866	0.864
10	C7	49	C	0.33458	No	0.000	0.000
11	C8	26	C	0.34843	No	0.000	0.000
12	C9	42	C	0.30096	No	0.000	0.000
13	L1	27	II	1.01316	Yes	1.000	1.000
14	L10	49	II	0.72741	Yes	0.693	0.676

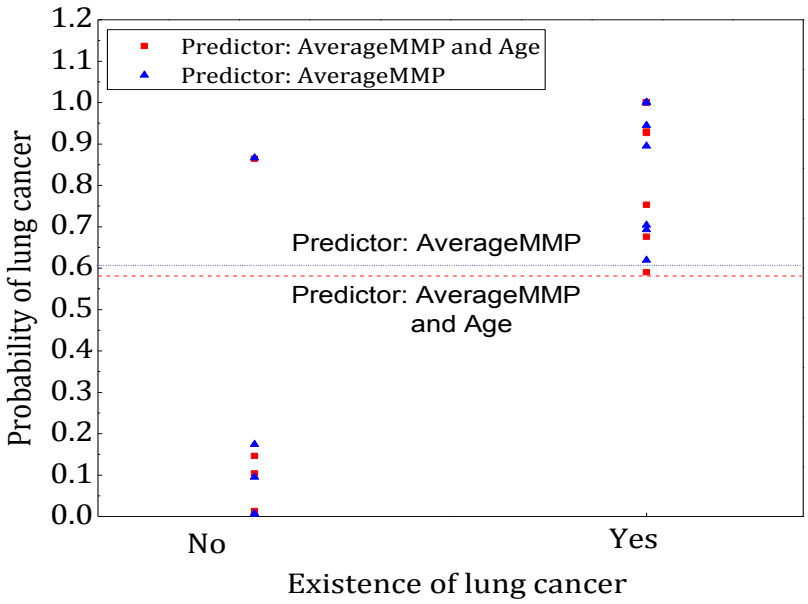
15	L11	40	III	0.72932	Yes	0.704	0.753
16	L12	50	II	0.71529	Yes	0.619	0.590
17	L2	44	III	1.04613	Yes	1.000	1.000
18	L3	63	II	1.00416	Yes	1.000	1.000
19	L4	54	III	1.03366	Yes	1.000	1.000
20	L5	37	II	0.77618	Yes	0.895	0.926
21	L6	46	III	1.07635	Yes	1.000	1.000
22	L7	51	I	1.04785	Yes	1.000	1.000
23	L8	38	II	0.97629	Yes	0.999	1.000
24	L9	55	III	0.80185	Yes	0.945	0.930

Table 5-10 summarizes the differences of the two models in the diagnostic test of lung cancer. Including age does not have much effect on the area under the ROC curve, optimal cut-off probability, sensitivity, specificity, or p-value for average MMP. The comparison is also shown in Figure 5-10.

Table 5-10 Comparison of models in section 5.2.1 and 5.2.2

Predictor	Area under the ROC curve	Threshold	Sensi	Speci	P-value	
					MMP	Age
AverageCath	0.9792	0.619	1	0.92	0.1146	N/A
AverageCath+Age	0.9792	0.589	1	0.92	0.1037	0.7574

Figure 5-10 Prediction of existence of cancer based on the optimal cut-off probability for the models used in section 5.2.1 and 5.2.2



5.4 Analysis based on multicategory logistic model

In this section, average MMP and age are used as the predictors to predict the probabilities of each stage of cancer for different persons. Analysis of effects of different parameters is shown in Table 5-11. Estimates of parameters are shown in Table 5-12. Table 5-12 lists the intercepts and coefficients of the three equations fit for the three cancer stages, I, II, and III. The intercept and coefficient of the equation for control group (including healthy persons and the patients with breast cancer of stage 0) are set to be 0 by default. Probabilities of each stage of lung cancer for different persons are computed and shown in Table 5-13. For example, the first line of Table 5-13 indicates the first patient has no lung cancer. Based on the diagnostic test, the probability that he or she has no lung cancer is 0.999. The probability that he or she is at stage I, II, or III is 0.000, 0.001, and 0.000, respectively. The p-values of average MMP and age are 0.3207 and 0.8878.

Table 5-11 Analysis of Effects

Effect	DF	Chi-Square	Pr > ChiSq
average MMP	3	3.5006	0.3207
Age	3	0.6803	0.8778

Table 5-12 Intercepts and coefficients of the logistic regression equations fit for the three cancer stages

Standard Parameter	Staging	DF	Estimate	Error	Chi-Square	Pr > ChiSq
Intercept	III	1	-22.4031	13.0225	2.9596	0.0854
Intercept	II	1	-16.9139	12.0692	1.964	0.1611
Intercept	I	1	-75.3616	104.6	0.5193	0.4711
average MMP	III	1	30.57	17.5026	3.0506	0.0807
average MMP	II	1	26.9478	17.0712	2.4918	0.1144
average MMP	I	1	74.639	86.9727	0.7365	0.3908
Age	III	1	-0.00405	0.128	0.001	0.9748
Age	II	1	-0.0484	0.1198	0.1634	0.686
Age	I	1	0.1129	0.352	0.1028	0.7485

Table 5-13 Probabilities of lung cancer in each stage for each patient

Obs	Age	Staging	averageMMP	Prob of no cancer	Prob of stage I	Prob of stage II	Prob of stage III
1	47	C	0.45603	0.999	0.000	0.001	0.000
2	51	C	0.32572	1.000	0.000	0.000	0.000
3	28	C	0.51654	0.986	0.000	0.013	0.001
4	32	C	0.61482	0.852	0.000	0.128	0.020
5	62	C	0.64015	0.897	0.000	0.062	0.041
6	48	C	0.76616	0.136	0.000	0.555	0.309
7	49	C	0.33458	1.000	0.000	0.000	0.000
8	26	C	0.34843	1.000	0.000	0.000	0.000
9	42	C	0.30096	1.000	0.000	0.000	0.000
10	60	C	0.29954	1.000	0.000	0.000	0.000
11	26	C	0.36683	1.000	0.000	0.000	0.000
12	44	C	0.41336	1.000	0.000	0.000	0.000
13	27	II	1.01316	0.000	0.002	0.649	0.349
14	44	III	1.04613	0.000	0.092	0.397	0.511
15	63	II	1.00416	0.000	0.159	0.236	0.605
16	54	III	1.03366	0.000	0.181	0.281	0.538
17	37	II	0.77618	0.075	0.000	0.683	0.242
18	46	III	1.07635	0.000	0.346	0.255	0.400
19	51	I	1.04785	0.000	0.220	0.282	0.498
20	38	II	0.97629	0.000	0.002	0.565	0.433
21	55	III	0.80185	0.066	0.000	0.501	0.433
22	49	II	0.72741	0.326	0.000	0.448	0.226
23	40	III	0.72932	0.250	0.000	0.559	0.191
24	50	II	0.71529	0.413	0.000	0.390	0.197

For each person, the stage of cancer is predicted to be the one with the largest probability. As a summary, table 5-14 lists the numbers of persons predicted to be lung cancers of different stages.

Table 5-14 Prediction of staging of lung cancer for patients in different stages of lung cancer

	Predicted Staging			
Staging	C	II	III	Sum
C	11	1		12
I			1	1
II	1	4	1	6
III		2	3	5
Sum	12	7	5	24

5.5 Analysis based on CART

The data are also analyzed by a classification tree. These patients are divided into two groups based on the method used in section 5.3. If the patient has no lung cancer, he is in the no lung cancer group. If the patient has lung cancer at stage I, II, or III, he is in the lung cancer group. The output is shown in Figure 5-11. If the average MMP is greater than 0.6777, the patient is predicted to have lung cancer. On the contrary, if the average MMP of the individual is smaller than 0.6777, the individual is predicted to have no lung cancer. Only one person without lung cancer is predicted to have lung cancer. The prediction is correct for the other persons. The prediction for each patient is in Table 5-16.

Figure 5-11 Predicted lung cancer condition by tree model

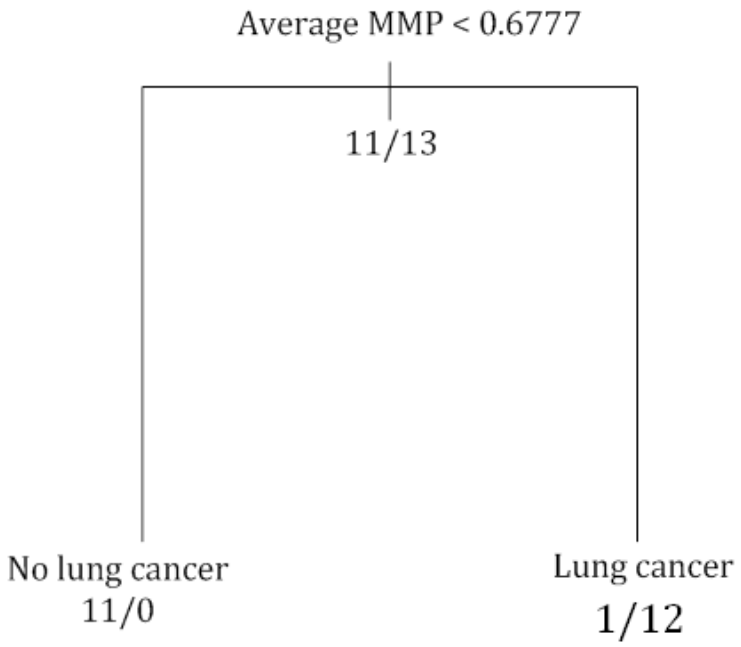


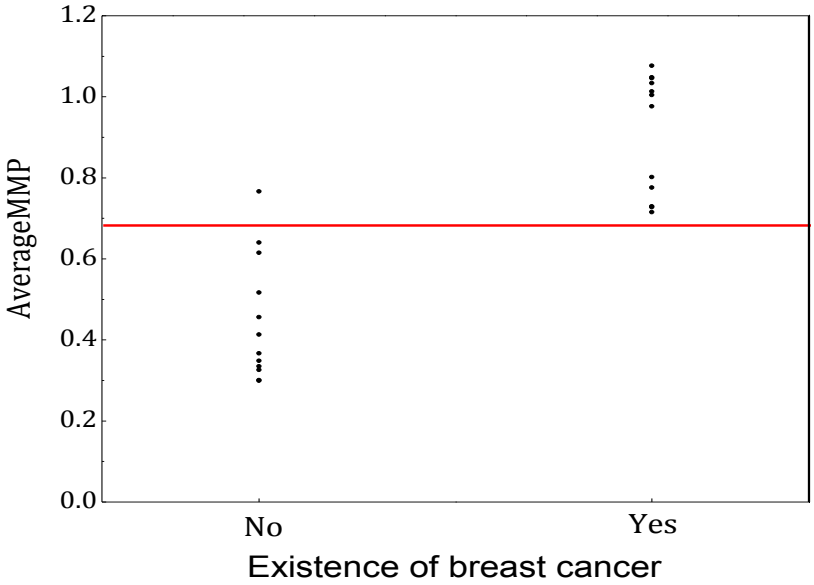
Table 5-15 Predicted lung cancer condition by tree model

Patient	Age	Staging	MMP1(1)	MMP1(2)	MMP1(3)	Average MMP	Cancer
C1	47	C	0.430	0.460	0.478	0.456	No
C2	51	C	0.308	0.329	0.340	0.326	No
C3	28	C	0.515	0.505	0.530	0.517	No
C4	32	C	0.612	0.609	0.624	0.615	No
C5	62	C	0.624	0.651	0.645	0.640	No
C6	48	C	0.769	0.756	0.773	0.766	Yes

C7	49	C	0.308	0.357	0.338	0.335	No
C8	26	C	0.356	0.301	0.389	0.348	No
C9	42	C	0.297	0.305	0.300	0.301	No
C10	60	C	0.289	0.310	0.300	0.300	No
C11	26	C	0.356	0.378	0.367	0.367	No
C12	44	C	0.404	0.413	0.423	0.413	No
L1	27	II	0.977	1.094	0.968	1.013	Yes
L2	44	III	1.084	1.055	1.000	1.046	Yes
L3	63	II	1.009	1.027	0.976	1.004	Yes
L4	54	III	0.997	1.057	1.047	1.034	Yes
L5	37	II	0.799	0.756	0.773	0.776	Yes
L6	46	III	1.138	1.063	1.028	1.076	Yes
L7	51	I	1.067	1.048	1.028	1.048	Yes
L8	38	II	1.007	0.967	0.955	0.976	Yes
L9	55	III	0.813	0.789	0.803	0.802	Yes
L10	49	II	0.727	0.753	0.703	0.727	Yes
L11	40	III	0.691	0.767	0.729	0.729	Yes
L12	50	II	0.689	0.714	0.743	0.715	Yes

The prediction for patients with or without lung cancer is illustrated in Figure 5-12. It can be seen that no patient with lung cancer is predicted to have lung cancer, whereas, one person without lung cancer is predicted to have lung cancer.

Figure 5-12 Scatter plot of enzyme pattern vs existence of lung cancer



It needs to be noted that the number of wrongly predicted person is the same as the number of wrongly predicted persons in the logistic regression model when average MMP is used as the predictor and this grouping is used (model in section 5.2.1).

Chapter 6 - Conclusion

This report illustrates how to use the method of logistic regression or the method of CART to investigate the performance of a new method developed by Dr. Bossmann and his coworkers in Kansas State University. It is found that the performance of these diagnostic tests is very good on these particular datasets, given that the area under the ROC curve is greater than 0.9. The nanoparticles developed by Dr. Bossmann's group appear to show promise in detecting breast cancer and lung cancer. Including the factor of patient's age helps to predict the existence of breast cancer for the samples used in this report. But since the samples sizes are small, it is impossible to conclude that including the age variable in the model helps the prediction of breast cancer. Logistic regression gives a better prediction than CART when age is included as one of the predictors. But the ability of these particles to predict the actual stages of cancer is low.

The results of this study have some limitations. First, data are obtained by observational study, and the persons involved in this study are not randomly chosen. Second, sample sizes are small. Third, though the performance of the test seems good, there are still false positives and negatives for some patients.

REFERENCES

1. Altman, D.G. and Bland, J.M. Diagnostic tests 1: sensitivity and specificity. BMJ. 1994, Vol. 308, p. 1552.

2. EBM notebook. On some clinically useful measures of the accuracy of diagnostic tests. Evidence-Based Medicine. 1998, Vol. 3.

3. Altman, D.G. and Bland, J.M. Diagnostic tests 2: predictive values. BMJ. 1994, 309.

4. Deeks, J.J. and Altman, D.G. Diagnostic tests 4: likelihood ratios. BMJ. 2004, Vol. 329.

5. Agresti, A. An introduction to categorical data analysis. s.l. : A John Wiley & Sons, Inc., Publication, 2007.

6. Cougmlin, S.S., et al. The logistic modeling of sensitivity, specificity, and predictive value of a diagnostic test. Jounal of Clinical Epidemiology. 1992, Vol. 45.

7. Hosmer, D.W. and Lemeshow, S. Applied Logistic Regression. s.l. : John WIley & Sons, Inc., 2000.

8. Flach, P. ROC Analysis. Encyclopedia of Machine Learning. s.l. : Springer.

9. MedicalBiostatistics.com. ROC Curve. [Online]

10. Hanley, J.A. and McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982, Vol. 143.

11. Yohannes, Y. and Webb, P. Classification and Regression Trees, CART. Wahington, D.C. : International Food Policy Research Institute, 1999.

12. Breiman, L., et al. Classification and Regression Trees. s.l. : Chapman & Hall/CRC, 1998.

13. Hastie, T.J. and Chambers, J.M. Statistical Models in S. s.l. : Chapman & Hall, Inc., 1993.

Appendix A - SAS and R codes

SAS code for section 3.2

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
              getnames=yes;
run;
proc print data=breast;
run;
data breastnew;
    set breast;
       if Staging="C" then level="moderate";
       else if Staging="0" then level="moderate";
       else if Staging="I" then level="moderate";
       else if Staging="II" then level="moderate";
       else if Staging="III" then level="severe";
       else if Staging="IV" then level="severe";
run;
proc logistic descending;
  model level=CathB1 Age;
run;
data probability1;
  set breastnew(keep=Patient CathB1 Age staging level);
  alpha1=-10.3277;
  beta1=3.336*0.000001;
  gamma1=0.0953;
  probability1=exp(alpha1+beta1*CathB1+gamma1*Age)/(1+exp(alpha1+beta1*CathB1+gamma1*Age));
run;
proc logistic descending;
  model level=CathB2 Age;
run;
data probability2;
  set breastnew(keep=Patient CathB2 Age staging level);
  alpha2=-10.1293;
  beta2=3.389*0.000001;
  gamma2=0.0922;
      probability2=exp(alpha2+beta2*CathB2+gamma2*Age)/(1+exp(alpha2+beta2*CathB2+gamma2*Age));
run;
proc logistic descending;
  model level=CathB3 Age;
run;
data probability3;
  set breastnew(keep=Patient CathB3 Age staging level);
  alpha3=-9.6663;
  beta3=3.407*0.000001;
  gamma3=0.0837;
probability3=exp(alpha3+beta3*CathB3+gamma3*Age)/(1+exp(alpha3+beta3*CathB3+gamma3*Age));
run;
proc sort data=probability1;
    by Patient;
run;
proc sort data=probability2;
```

```
    by Patient;
run;
proc sort data=probability3;
    by Patient;
run;
data comparison;
        merge probability1 probability2 probability3;
        by Patient;
        keep Patient  Age Staging level probability1 probability2 probability3;
run;
proc print data=comparison;
run;
ODS CSV file="D:\master report\comparison of three probabilities.csv";
proc print data=comparison;
run;
ODS CSV close;
run;
```

SAS code for section 3.3.1

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
          getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then Cancer="moderate";
       else if Staging="0" then delete;
       else if Staging="I" then delete;
       else if Staging="II" then Cancer="severe";
       else if Staging="III" then Cancer="severe";
       else if Staging="IV" then Cancer="severe";
run;
proc logistic descending;
  model Cancer=CathB/outroc=roc;
run;
data probability;
  set breastnew(keep=Patient CathB Age staging Cancer);
  probability=exp(-2.8809+0.000002532*CathB)/(1+exp(-2.8809+0.000002532*CathB));
ODS CSV file="D:\master report\master report\ averageCathB as predictors for CathB (stage 0 and I
deleted).csv";
proc print data=probability;
run;
ODS CSV close;
run;

ODS CSV file="D:\master report\ROC\ROC3.4.1.csv";
proc print data=roc;
run;
ODS CSV close;
run;
ods graphics on;
ods html;
proc logistic data=breastnew desc plots(only)=(roc);
model Cancer =CathB;
run;
ods html close;
ods graphics off;
run;
```

SAS code for section 3.3.2

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
          getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then Cancer ="moderate";
       else if Staging="0" then delete;
       else if Staging="I" then delete;
       else if Staging="II" then Cancer ="severe";
       else if Staging="III" then Cancer ="severe";
       else if Staging="IV" then Cancer ="severe";
run;
proc logistic descending;
  model Cancer =CathB Age/outroc=roc;
run;
data probability;
  set breastnew(keep=Patient CathB Age staging Cancer);
probability=exp(-6.2938+0.000002073*CathB+0.0785*Age)/(1+exp(-
6.2938+0.000002073*CathB+0.0785*Age));
ODS CSV file="D:\master report\master report\ averageCathB and age as predictors for CathB (stage 0 and I
deleted).csv";
proc print data=probability;
run;
ODS CSV close;
run;

ODS CSV file="D:\master report\ROC\ROC3.4.csv";
proc print data=roc;
run;
ODS CSV close;
run;
ods graphics on;
ods html;
proc logistic data=breastnew desc plots(only)=(roc);
model Cancer =CathB Age;
run;
ods html close;
ods graphics off;
run;
```

SAS code for section 3.4.1

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
               getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then level="moderate";
       else if Staging="0" then level="moderate";
       else if Staging="I" then level="moderate";
       else if Staging="II" then delete;
       else if Staging="III" then level="severe";
       else if Staging="IV" then level="severe";
run;
proc logistic descending;
  model level=CathB/covb outroc=roc;
run;
data probability;
  set breastnew(keep=Patient CathB staging level);
  alpha=-4.8059;
  beta=3.381*0.000001;
  probability=exp(alpha+beta*CathB)/(1+exp(alpha+beta*CathB));
  varalpha=3.060474;
  varbeta=1.59*0.000000000001;
  cov=-2.05*0.000001;
  SE=sqrt(varalpha+CathB*CathB*varbeta+2*CathB*cov);
  lower=alpha+beta*CathB-1.96*SE;
  upper=alpha+beta*CathB+1.96*SE;
  lowerpro=exp(lower)/(1+exp(lower));
  upperpro=exp(upper)/(1+exp(upper));
run;
proc print data=probability (keep=Patient Staging level probability lowerpro upperpro);
run;
ODS CSV file="D:\master report\master report\ averageCathB as predictors for CathB (stage II deleted).csv";
proc print data=probability;
run;
ODS CSV close;
run;
ODS CSV file="D:\master report\ROC\ROC3.3.1.csv";
proc print data=roc;
run;
ODS CSV close;
run;
ods graphics on;
ods html;
      proc logistic data=breastnew desc plots(only)=(roc);
      model level=CathB;
run;
ods html close;
ods graphics off;
run;
```

SAS code for section 3.4.2

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
               getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then level="moderate";
       else if Staging="0" then level="moderate";
       else if Staging="I" then level="moderate";
       else if Staging="II" then delete;
       else if Staging="III" then level="severe";
       else if Staging="IV" then level="severe";
run;
proc logistic descending;
  model level=CathB Age/outroc=roc;
run;
data probability2;
  set breastnew(keep=Patient CathB Age staging level);
  probability2=exp(-9.4669+3.254*0.000001*CathB+0.0879*Age)/(1+exp(-
9.4669+3.254*0.000001*CathB+0.0879*Age));
proc print data=probability2;
run;
ODS CSV file="D:\master report\master report\ averageMMP and age as predictors for CathB.csv";
proc print data=probability2;
run;
ODS CSV close;
run;

ODS CSV file="D:\master report\ROC\ROC3.3.2.csv";
proc print data=roc;
run;
ODS CSV close;
run;
ods graphics on;
ods html;
      proc logistic data=breastnew desc plots(only)=(roc);
      model level=CathB Age;
run;
ods html close;
ods graphics off;
run;
```

SAS code for section 3.4.3

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
              getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then level="moderate";
       else if Staging="0" then level="moderate";
       else if Staging="I" then level="moderate";
       else if Staging="II" then delete;
       else if Staging="III" then level="severe";
       else if Staging="IV" then level="severe";
run;
proc logistic descending;
  model level=CathB/covb;
run;
data probability;
  set breastnew(keep=Patient CathB staging level);
  alpha=-4.8059;
  beta=3.381*0.000001;
  probability=exp(alpha+beta*CathB)/(1+exp(alpha+beta*CathB));
run;
proc logistic descending;
  model level=CathB Age;
run;
data probability2;
  set breastnew(keep=Patient CathB Age staging level);
probability2=exp(-9.4669+3.254*0.000001*CathB+0.0879*Age)/(1+exp(-
9.4669+3.254*0.000001*CathB+0.0879*Age));
proc sort data=probability;
    by Patient;
run;
proc sort data=probability2;
    by Patient;
run;
data comparison;
         merge probability probability2;
         by Patient;
         keep Patient CathB Age Staging level probability probability2;
run;
ODS CSV file="D:\master report\comparison(Stage II delete).csv";
proc print data=comparison;
run;
ODS CSV close;
run;
```

SAS code for section 3.5.1

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
               getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then level="moderate";
       else if Staging="0" then level="moderate";
       else if Staging="I" then level="moderate";
       else if Staging="II" then level="moderate";
       else if Staging="III" then level="severe";
       else if Staging="IV" then level="severe";
run;
proc logistic descending;
  model level=CathB/outroc=roc;
run;
data probability;
  set breastnew(keep=Patient CathB staging level);
probability=exp(-5.2329+3.535*0.000001*CathB)/(1+exp(-5.2329+3.535*0.000001*CathB));
proc print data=probability;
run;
ODS CSV file="D:\master report\master report\ averageCathB as predictors for CathB (complete data).csv";
proc print data=probability;
run;
ODS CSV close;
run;
          ODS CSV file="D:\master report\ROC\ROC3.4.1.csv";
          proc print data=roc;
run;
ODS CSV close;
ods graphics on;
ods html;
          proc logistic data=probability desc plots(only)=(roc);
          model level=CathB;
run;
ods html close;
ods graphics off;
```

SAS code for section 3.5.2

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
               getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then level="moderate";
       else if Staging="0" then level="moderate";
       else if Staging="I" then level="moderate";
       else if Staging="II" then level="moderate";
       else if Staging="III" then level="severe";
       else if Staging="IV" then level="severe";
run;
proc logistic descending;
  model level=CathB Age/outroc=roc;
run;
data probability2;
  set breastnew(keep=Patient CathB Age staging level);
probability2=exp(-9.9959+3.374*0.000001*CathB+0.0898*Age)/(1+exp(-
9.9959+3.374*0.000001*CathB+0.0898*Age));
proc print data=probability2;
run;
ODS CSV file="D:\master report\averageCathB and age as predictor(complete data).csv";
proc print data=probability2;
run;
ODS CSV close;
ODS CSV file="D:\master report\ROC\ROC3.4.2.csv";
proc print data=roc;
run;
ODS CSV close;
run;
ods graphics on;
ods html;
          proc logistic data=breastnew plots(only)=(roc);
          model level=CathB Age;
run;
ods html close;
ods graphics off;
```

SAS code for section 3.5.3

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
               getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
    if Staging="C" then level="moderate";
       else if Staging="0" then level="moderate";
       else if Staging="I" then level="moderate";
       else if Staging="II" then level="moderate";
       else if Staging="III" then level="severe";
       else if Staging="IV" then level="severe";
run;
proc logistic descending;
  model level=CathB;
run;
data probability;
  set breastnew(keep=Patient CathB staging level);
probability=exp(-5.2329+3.535*0.000001*CathB)/(1+exp(-5.2329+3.535*0.000001*CathB));
proc logistic descending;
  model level=CathB Age;
run;
data probability2;
  set breastnew(keep=Patient CathB Age staging level);
probability2=exp(-9.9959+3.374*0.000001*CathB+0.0898*Age)/(1+exp(-
9.9959+3.374*0.000001*CathB+0.0898*Age));
proc sort data=probability;
    by Patient;
run;
proc sort data=probability2;
    by Patient;
run;
data comparison;
         merge probability probability2;
         by Patient;
         keep Patient CathB Age Staging level probability probability2;
run;
proc print data=comparison;
run;
```

SAS code for section 3.6

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\breast cancer.xls"
     dbms=xls replace;
          getnames=yes;
run;
data breastnew;
    set breast;
    CathB=1/3*(CathB1 +CathB2+CathB3);
          if Staging="0" then Staging="C";
run;
proc logistic descending;
  model Staging=CathB Age/link=glogit;
run;
data probability;
  set breastnew(keep=Patient CathB Staging Age);
  alpha1=-12.3027; beta1=-0.00000334;  gamma1=0.2163;
  alpha2=-4.7017;  beta2= 0.0000005452; gamma2=0.0581;
  alpha3=-11.3583; beta3= 0.000002923;  gamma3=0.1295;
  alpha4=-14.1369; beta4= 0.000003849;  gamma4= 0.1430;
  sum=exp(alpha1+beta1*CathB+gamma1*Age)+exp(alpha2+beta2*CathB+gamma2*Age)+
  exp(alpha3+beta3*CathB+gamma3*Age)+exp(alpha4+beta4*CathB+gamma4*Age)+1;
  PI=exp(alpha1+beta1*CathB+gamma1*Age)/sum;
  PII=exp(alpha2+beta2*CathB+gamma2*Age)/sum;
  PIII=exp(alpha3+beta3*CathB+gamma3*Age)/sum;
  PIV=exp(alpha4+beta4*CathB+gamma4*Age)/sum;
  PC=1-PI-PII-PIII-PIV;
run;
ODS CSV file="D:\master report\multi-stages probability for CathB.csv";
proc print data=probability (keep = CathB Staging Age PI PII PIII PIV PC);
run;
ODS CSV close;
run;
```

R code for section3.8

```
# install rpart library
      chooseCRANmirror()
      install.packages("rpart")
library(rpart)

#read data
CathB=read.csv("D:master report\\CathB-revised.csv")
B1=CathB[1:32,] #These are the breast cancer and control data
averageCath=1/3*(B1[,5]+B1[,6]+B1[,7])
B2=cbind(B1, averageCath)

#tree classification
fit <- rpart(Staging ~ averageCath + Age,
            method="class", data=B1)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# plot tree
plot(fit, uniform=TRUE,
            main="Classification Tree for CathB")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

SAS code for section 4.2

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\master report\breast cancer\MP1breast.csv"
     dbms=csv replace;
               getnames=yes;
run;
proc print data=breast;
run;
data breastnew;
    set breast;
       if Staging="C" then Cancer="No";
       else if Staging="0" then Cancer="No";
       else if Staging="I" then Cancer ="Yes";
       else if Staging="II" then Cancer ="Yes";
       else if Staging="III" then Cancer ="Yes";
       else if Staging="IV" then Cancer ="Yes";
run;
proc logistic descending;
  model Cancer=MMP11 Age;
run;
data probability1;
  set breastnew(keep=Patient MMP11 Age staging Cancer);
  alpha1=-10.3277;
  beta1=3.336*0.000001;
  gamma1=0.0953;
  probability1=exp(alpha1+beta1*MMP111+gamma1*Age)/(1+exp(alpha1+beta1*MMP11+gamma1*Age));
run;
proc logistic descending;
  model Cancer=MMP12 Age;
run;
data probability2;
  set breastnew(keep=Patient MMP12 Age staging Cancer);
  alpha2=-10.1293;
  beta2=3.389*0.000001;
  gamma2=0.0922;
      probability2=exp(alpha2+beta2*MMP12+gamma2*Age)/(1+exp(alpha2+beta2*MMP12+gamma2*Age));
run;
proc logistic descending;
  model Cancer=MMP13 Age;
run;
data probability3;
  set breastnew(keep=Patient MMP13 Age staging Cancer);
  alpha3=-9.6663;
  beta3=3.407*0.000001;
  gamma3=0.0837;
probability3=exp(alpha3+beta3*MMP13+gamma3*Age)/(1+exp(alpha3+beta3*MMP13+gamma3*Age));
run;
proc sort data=probability1;
    by Patient;
run;
proc sort data=probability2;
    by Patient;
run;
proc sort data=probability3;
```

```
    by Patient;
run;
data comparison;
        merge probability1 probability2 probability3;
        by Patient;
        keep Patient  Age Staging Cancer probability1 probability2 probability3;
run;
proc print data=comparison;
run;
ODS CSV file="D:\master report\comparison of three probabilities for breast cancer (MP1).csv";
proc print data=comparison;
run;
ODS CSV close;
run;
```

SAS code for section 4.3.1

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\master report\breast cancer\MP1breast.csv"
     dbms=csv replace;
              getnames=yes;
run;
data breastnew;
    set breast;
    MMP=1/3*(MMP11 +MMP12+MMP13);
    if Staging="C" then Cancer="No";
       else if Staging="0" then Cancer="No";
       else if Staging="I" then Cancer ="Yes";
       else if Staging="II" then Cancer ="Yes";
       else if Staging="III" then Cancer ="Yes";
       else if Staging="IV" then Cancer ="Yes";
run;
proc logistic descending;
  model Cancer =MMP/outroc=roc;
run;
data probability;
  set breastnew(keep=Patient MMP Age staging Cancer);
probability=exp(-6.5605+11.5520*MMP)/(1+exp(-6.5605+11.5520*MMP));
proc print data=probability;
run;
ODS CSV file="D:\master report\master report\breast cancer\averageCathB as predictor(MP1).csv";
proc print data=probability;
run;
ODS CSV close;
ODS CSV file="D:\master report\ROC\ROC5.3.1.csv";
proc print data=roc;
run;
ODS CSV close;
ods graphics on;
ods html;
proc logistic data=breastnew plots(only)=(roc);
   model Cancer =MMP;
   run;
ods html close;
```

SAS code for section 4.3.2

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\master report\breast cancer\MP1breast.csv"
     dbms=csv replace;
              getnames=yes;
run;
data breastnew;
    set breast;
    MMP=1/3*(MMP11 +MMP12+MMP13);
    if Staging="C" then Cancer="No";
       else if Staging="0" then Cancer="No";
       else if Staging="I" then Cancer ="Yes";
       else if Staging="II" then Cancer ="Yes";
       else if Staging="III" then Cancer ="Yes";
       else if Staging="IV" then Cancer ="Yes";
run;
proc logistic descending;
  model Cancer=MMP Age/outroc=roc;
run;
data probability;
  set breastnew(keep=Patient MMP Age staging Cancer);
probability=exp(-10.5871+10.2277*MMP+0.0970*Age)/(1+exp(-10.5871+10.2277*MMP+0.0970*Age));
proc print data=probability;
run;
ODS CSV file="D:\master report\master report\breast cancer\averageCathB and age as predictor(MP1).csv";
proc print data=probability;
run;
ODS CSV close;
ODS CSV file="D:\master report\ROC\ROC5.3.2.csv";
          proc print data=roc;
          run;
ODS CSV close;
ods graphics on;
ods html;
proc logistic data=breastnew plots(only)=(roc);
   model Cancer=MMP Age;
   run;
ods html close;
ods graphics off;
```

SAS code for section 4.4

```
options nodate nonumber;
proc import out =work.breast datafile="D:\master report\master report\breast cancer\MP1breast.csv"
     dbms=csv replace;
          getnames=yes;
run;
data breastnew;
    set breast;
    averageMMP=1/3*(MMP11 +MMP12+MMP13);
    if Staging="0" then Staging="C";
run;
proc logistic descending;
  model Staging=averageMMP Age/link=glogit;
run;
data probability;
  set breastnew(keep=Patient averageMMP Age staging);
  alpha1=-17.8201; beta1=10.8368;   gamma1=0.1704;
  alpha2=-9.1344;  beta2= 8.9862;   gamma2=0.0552;
  alpha3=-13.4324; beta3= 11.6952;  gamma3=0.1094;
  alpha4=-12.1065; beta4= 10.1164;  gamma4=0.1101;
  sum=exp(alpha1+beta1*averageMMP+gamma1*Age)+exp(alpha2+beta2*averageMMP+gamma2*Age)+
  exp(alpha3+beta3*averageMMP+gamma3*Age)+exp(alpha4+beta4*averageMMP+gamma4*Age)+1;
  PI=exp(alpha1+beta1*averageMMP+gamma1*Age)/sum;
  PII=exp(alpha2+beta2*averageMMP+gamma2*Age)/sum;
  PIII=exp(alpha3+beta3*averageMMP+gamma3*Age)/sum;
  PIV=exp(alpha4+beta4*averageMMP+gamma4*Age)/sum;
  PC=1-PI-PII-PIII-PIV;
run;

ODS CSV file="D:\master report\master report\breast cancer\averageMMP and age as predictor for
multicategory response.csv";
proc print data=probability (keep=averageMMP Staging Age PI PII PIII PIV PC);
run;
ODS CSV close;
run;
```

R code for section 4.5

```
library(rpart)
MP1=read.csv("D:master report\\MP1-revised.csv")
B1=MP1[1:32,] #These are the breast cancer and control data
averageMMP=1/3*(B1[,5]+B1[,6]+B1[,7])
B2=cbind(B1, averageMMP)

#tree classification
fit <- rpart(Staging ~ averageMMP + Age,
        method="class", data=B1)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# plot tree
plot(fit, uniform=TRUE,
        main="Classification Tree for MMPbreast")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

SAS code for section 5.3.1

```
proc import out =work.lung datafile="D:\master report\master report\lung cancer\MP1lungcancer.csv"
     dbms=csv replace;
             getnames=yes;
run;
data lungnew;
    set lung;
    averageMMP=1/3*(MMP11 +MMP12+MMP13);
    if Staging="C" then Cancer="No";
       else if Staging="I" then Cancer="Yes";
       else if Staging="II" then Cancer="Yes";
       else if Staging="III" then Cancer="Yes";
run;
proc logistic descending;
  model Cancer=averageMMP/outroc=roc;
run;
data probability;
  set lungnew(keep=Patient averageMMP staging Cancer);
      probability=exp(-18.9824+27.2166*averageMMP)/(1+exp(-18.9824+27.2166*averageMMP));
ODS CSV file="D:\master report\master report\lung cancer\averageMMP as predictor.csv";
          proc print data=probability;
          run;
ODS CSV close;
ODS CSV file="D:\master report\ROC\ROC4.3.1.csv";
          proc print data=roc;
          run;
ODS CSV close;

ods graphics on;
ods html;
proc logistic data=lungnew desc plots(only)=(roc);
   model Cancer=averageMMP;
   run;
ods html close;
ods graphics off;
run;
```

SAS code for section 5.3.2

```
options nodate nonumber;
proc import out =work.lung datafile="D:\master report\master report\lung cancer\MP1lungcancer.csv"
     dbms=csv replace;
              getnames=yes;
run;
data lungnew;
    set lung;
    averageMMP=1/3*(MMP11 +MMP12+MMP13);
    if Staging="C" then Cancer="No";
       else if Staging="I" then Cancer="Yes";
       else if Staging="II" then Cancer="Yes";
       else if Staging="III" then Cancer="Yes";
run;
proc logistic descending;
  model Cancer=averageMMP Age/outroc=roc;
run;
data probability2;
  set lungnew(keep=Patient averageMMP Age staging Cancer);
probability2=exp(-17.6394+27.7137*averageMMP-0.0364*Age)/(1+exp(-
17.6394+27.7137*averageMMP-0.0364*Age));
ODS CSV file="D:\master report\master report\lung cancer\averageMMP and Age as predictor.csv";
proc print data=probability2;
run;
ODS CSV close;
ODS CSV file="D:\master report\ROC\ROC4.3.2.csv";
          proc print data=roc;
          run;
ODS CSV close;
ods graphics on;
ods html;
proc logistic data=lungnew plots(only)=(roc);
   model Cancer=averageMMP Age;
   run;
ods html close;
ods graphics off;
run;
```

SAS code for section 5.3.3

```
proc import out =work.lung datafile="D:\master report\master report\lung cancer\MP1lungcancer.csv"
     dbms=csv replace;
          getnames=yes;
run;
data lungnew;
    set lung;
    averageMMP=1/3*(MMP11 +MMP12+MMP13);
    if Staging="C" then Cancer="No";
      else if Staging="I" then Cancer="Yes";
      else if Staging="II" then Cancer="Yes";
      else if Staging="III" then Cancer="Yes";
run;
proc logistic descending;
  model Cancer=averageMMP;
run;
data probability;
  set lungnew(keep=Patient averageMMP staging Cancer);
probability=exp(-18.9824+27.2166*averageMMP)/(1+exp(-18.9824+27.2166*averageMMP));
run;
proc logistic descending;
  model Cancer=averageMMP Age;
run;
data probability2;
  set lungnew(keep=Patient averageMMP Age staging Cancer);
probability2=exp(-17.6394+27.7137*averageMMP-0.0364*Age)/(1+exp(-
17.6394+27.7137*averageMMP-0.0364*Age));
run;
proc sort data=probability;
    by Patient;
run;
proc sort data=probability2;
    by Patient;
run;
data comparison;
         merge probability probability2;
         by Patient;
         keep Patient averageMMP Age Staging Cancer probability probability2;
run;
ODS CSV file="D:\master report\master report\lung cancer\comparison.csv";
proc print data=comparison;
run;
ODS CSV close;
run;
```

SAS code for section 5.4

```
options nodate nonumber;
proc import out =work.lung datafile="D:\master report\master report\lung cancer\MP1lungcancer.csv"
     dbms=csv replace;
          getnames=yes;
run;
data lungnew;
    set lung;
    averageMMP=1/3*(MMP11 +MMP12+MMP13);
run;
proc logistic descending;
  model Staging=averageMMP Age/link=glogit;
run;
data probability;
  set lungnew(keep=Patient averageMMP Age Staging);
  alpha1=-75.3616; beta1=74.6390;   gamma1=0.1129;
  alpha2=-16.9139;  beta2= 26.9478; gamma2=-0.0484;
  alpha3=-22.4031; beta3= 30.5700;  gamma3=-0.00405;
  sum=exp(alpha1+beta1*averageMMP+gamma1*Age)+exp(alpha2+beta2*averageMMP+gamma2*Age)+
  exp(alpha3+beta3*averageMMP+gamma3*Age)+1;
  PI=exp(alpha1+beta1*averageMMP+gamma1*Age)/sum;
  PII=exp(alpha2+beta2*averageMMP+gamma2*Age)/sum;
  PIII=exp(alpha3+beta3*averageMMP+gamma3*Age)/sum;
  PC=1-PI-PII-PIII;
run;
ODS CSV file="D:\master report\master report\lung cancer\averageMMP and Age as predictor
(multicategory response).csv";
proc print data=probability (keep=averageMMP Age Staging PC PI PII PIII);
run;
ODS CSV close;
run;
```

R code for section 5.5

```
library(rpart)

#read data
MP=read.csv("D:master report\\MP1-revised.csv")
L1=MP[21:44,] #These are the lung cancer and control data
averageMMP1=1/3*(MP[21:44,5]+MP[21:44,6]+MP[21:44,7])
L2=cbind(L1, averageMMP1)
attach(L2)

#tree classification
fit <- rpart(Staging ~ averageMMP1 + Age,
        method="class", data=L2)
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# plot tree
plot(fit, uniform=TRUE,
        main="Classification Tree for lung cancer")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```