International Journal of IT-based Public Health Management
Vol. 3, No. 2 (2016), pp. 15-20
http://dx.doi.org/10.21742/ijiphm.2016.3.2.03

# Statistical Techniques for Breast Cancer Classification: A Review

Aashna Vijay and Geetika Munja[1]

*The Northcap University, Sector-23A, Gurgaon, Haryana*

***Abstract***

*Microarrays gene expressions have played a vital role in the prognosis of genetic diseases. DNA microarrays are used to measure expressions of several genes simultaneously. They are not only used to determine variations in gene sequences but also used to identify the drug treatment best suited to the various genetic diseases like cancer. This paper contains the introduction on DNA microarrays and various types of breast cancer. It also explains various statistical techniques which are used in Survival Analysis (SA). The SA is used to study the estimated time duration between one or more events happen. When information is incomplete about survival time then it is called censored. These statistical methods can incorporate both censored as well as uncensored data to analyze the survival time of patients. Kaplan- Meier graph, Log rank test and Cox regression analysis are few statistical tools to study survival analysis.*

***Keywords:*** *Microarrays gene expressions, DNA Microarrays, Kaplan-Meier graph, Cox regression analysis, etc.*

## 1. Introduction

Scientists already know that a mutation in DNA's can lead to various diseases. It's very difficult to detect mutations as there are several regions in large genes where alterations can take place. A tool used in order to determine whether DNA of an individual contains a mutation in genes is called DNA microarray (also commonly known as DNA chip or biochip) [1]. With the advent of DNA microarray technology, researchers are empowered to study the growth level and development stages of life as well as examine the genetic reasons of anomalies developing in human body. A DNA chip is made up of a glass plate embedded in plastic. Each chip consists of minute DNA spots attached to the surface. DNA spots contain thousands of single-stranded, short synthetic DNA strands that group to form regular genes and its alterations found in the human beings [2]].

Initially, mRNA molecule originates from DNA sequences. Later, microarray experiment requires an mRNA molecule to hybridize with the DNA sequences from which it has been developed. Samples of DNA sequences are used to fabricate an array. The expression level of several genes indicates the amount of mRNA stick to a point on the array. Data related to gene expression is consolidated and profiled. Thousands of spotted samples are present on a solid support (a silicon chip or a microscopic glass slide). Spots can be oligonucleotides, cDNA or DNA.

**Article history:**
Received (March 21, 2016),Review Result (May 07, 2016), Accepted (June 28, 2016)

[1]**Log-Rank Test:** The survival curves drawn using Kaplan Meier graph need to beanalyzed more deeply using this technique. This is a non-parametric test and is used to analyze censored or right skewed data [5]. Study of the result is based on the number of times an event has occurred. The formula for test statistic is:

ISSN: 2205-8508 (Print), eISSN: 2207-3965 (Online) IJIPHM

Earlier, microarrays were used in research work only. Many massive population studies aresupervised by scientists- for example, to analyze frequency of people with a specific mutation that eventually leads to breast cancer, or to determine variations in genes which are mostly concerned with peculiar diseases [1]. This crucial advancement has been achieved only because of microarray chips which are very similar to computer chips. Microarray chips containnumerous features representing large amount of human genome. Today, Microarrays are not only being used in research work but also in various clinical diagnostic tests for diagnosing diseases. As genes are responsible to handle the chemistry between the drugs, microarrays are also being used to determine the drug which suits a person suffering a particular disease.

In order to determine that whether a person possesses an alteration in DNA for a certaindisease scientist firstly extracts DNA sample from the patient's blood followed by control sample- that doesn't contain any alteration in the gene of interest. Researchers create single stranded DNA molecule by separating the two hybridized DNA strands and label them with fluorescent dye. Healthy cell DNA is dyed in green color and cancerous cell DNA is dyed with red color. Once, DNA strands for each color is disseminated over the chip, hybridization occurs. Hybridization is the process in which DNA strands fuse with each other. Dyed strands hybridize with the complementary synthetic stranded DNA grooved on the array. When microarray is scanned under the microarray scanner different colors are displayed on the computer screen. Red color symbolizes amount of cancerous gene have expressed, green symbolizes expression of healthy genes and yellow color signifies uniform level activity for both kind of genes.

## 2. Molecular classification of breast cancer

Breast cancer is a type of cancer that evolves from breast tissue. Symptoms of breast cancer are lump in the breast, fluid arousing from nipple, change in the shape of breast or a red patch on the skin. Breast is formed of glands called lobules that produce milk. Milk travels from lobules to nipples through thin tubes called ducts. Structure of breast is very complicated that also includes connective tissue, lymph nodes, fat and blood vessels.

Ductal carcinoma is the most prevalent kind of breast cancer which initially develops in theduct cells. However, breast cancer can also originate from the cells of the lobules, tissue, etc. Histological classification of Breast cancer includes in situ and invasive breast cancer. When cancer hasn't spread to the surrounding tissues from the point it has evolved it is called in situ cancer whereas invasive cancer is vice-versa. Molecular subtypes of breast cancer are luminal A, luminal B, triple negative and HER2.

**Table 1. Difference between molecular classifications of breast cancer**

| | **Luminal** | **Triplenegative** | **HER2** |
|---|---|---|---|
| **Treatment** | Luminal tumors canbe treatedwith hormona therapy. | Triplenegative cancercar belook upon with the combination of surgery, radiation therapy and chemotherapy. | HER2 cancer can be treatedwithantiHER2drugs liketrastu-zumab. |

| | | | |
|---|---|---|---|
| **Gene expression** | High expression of hormonereceptors. | 90% oftriple negative tumorslackestrogen and progesteronereceptors. | These cancers overproduce HER2genes. |
| **Grade** | Luminalarelowgrade tumors. | Triplenegative arevery aggressive andoften classifiedundergrade3. | A grade of0or1signifies HER2negativeandgrade+2 or3 signifiesHER2positive. |
| Clinical features | About 70% of breast cancer patients suffer from either luminal cancers. Luminal B tumors are more aggressive than luminal A. | Triple negative tumors are fast growing, aggressive and spread very | HER2 tumors are the most aggressive tumors. About 5% to 15% of breast cancer patients are classified under |

## 3. Survival analysis

The survival analysis is the study of the estimated time span up till one or many instances occur [3]. It is the study on survival of people suffering from various diseases over a period of time. Individuals are followed till time at which event of interest occurs [3]. Events can be deaths, marriage, accidents, etc. When information is incomplete about survival time then it is called censored. Information can be left or right censored. A patient whose lifetime is less than specified duration, the lifetime is known as left censored. However, a patient is said to be right censored if he does not experience the event of interest in the time span of study. Also, before the termination of the experiment observation time if a person drops out of the experiment and didn't experience the event his survival time is right censored. Survival and Hazard functions depend on time and they are very important from the prospect of analyzing data [7]. The probability of surviving minimum to time t is known as survival function (S (t)). The conditional probability of Hazard function (h(t)) is the probability that a person will confront an event within an interval ensuring that person has lived till the starting of the interval.

### 3.1. Significance of gene expression

Genes are the carriers of hereditary information database that is transferred from parents to offspring. DNA is a hereditary material that contains genetic information used in the functioning of living organisms. The process in which information from a gene is utilized to produce various gene products like proteins is called gene expression. Gene expression profiling is the analysis of the expression of numerous genes at once to develop a worldwide image of cellular function. The profile developed from gene expression profiling helps in various tasks like distinguishing actively dividing cells, also show the reaction of cells on different treatments.

Gene expressions play vital roles like cancer susceptibility, testing penicillin resistance, that any individual's sensitivity to cancer can be determined by analyzing the expression levels of cancerous genes. For example tumor suppressor genes generate protein that suppresses the growth of tumor if such genes are under-expressed the risk of getting cancer increases. Secondly, testing penicillin resistance means that if bacteria destroy the antibiotic before antibiotic can kill the bacteria then antibiotic become ineffective. For example, the antibiotic penicillin can be deactivated by bacteria that produce an enzyme penicillinase. Thirdly, gene expression is the strictly regulated process in the body. If the process is not carried out strictly then this eventually leads to severe ramifications like cancer. For example

 

if insulin expressions are regulated then this controls the blood glucose levels in the body. Fourthly, measuring viral genes expression can also help in the development of vaccines by understanding the viral mechanism.

### 3.2 Statistical evaluation metrics for gene data

**3.2.1. Correlation:** Correlation is a technique that determines how strongly two variables are related to each other. It is easy to calculate and understand. Value of correlation coefficient lies between -1 to 1 [4]. If the value is 0 or near to 0 then variables are weakly correlated. However, if values are 1 or -1 this means that variables are strongly correlated. Two fundamental types of correlation coefficient are:

Pearson's Product Moment Correlation Coefficient: It is a technique that depicts the linear correlation between two variables [4]. Pearson's correlation coefficient represented as p signifies population and represented as r signifies sample statistic. This technique is applied if data examined in study is normally distributed [4]. It gets influence by the outliers which may overemphasize or weaken the relationship [4]. Therefore, it is unsuitable if neither of the variables are distributed normally. In this study of binary classification of cancer in into er+ and er- Pearson's correlation coefficient was calculated in order to find the most correlated genes. Thereafter, genes were arranged in the descending order of their correlation coefficient. This step was performed for feature selection. Formula of the technique is shown below:

$$\frac{\sum (p\ i \sim P)(Ri \sim R)}{\sum_{v,q} V(Pi-p)^2 Vfe-?)^2} \quad \textbf{(1)}$$

Here, p**i**andq**i**arethe valuesofgenes p and q forthei**th**sample.

**3.2.2. Spearman's rank correlation coefficient:** This technique is applicable when either one or both variables are skewed and outliers are present [4]. Its coefficient represented as ps denotes population parameter and as rs denotes sample statistic. The sample Spearman's correlation coefficient can be calculated between the variables p and q by:

$$6£d \setminus \textbf{(2)}$$

$$n(n2 — 1)\ '$$

Here di is the subtraction between the ranks of p and q.

**3.2.3. Kaplan-Meier analysis:** A Kaplan-Meier analysis is used to estimate the survival of people suffering from various health risks over the time. A Kaplan-Meier graph consists of two axis and two curves [3]. The probability or likelihood of survival is shown on vertical axis and time span in days, months or years is shown on horizontal axis. Typically each curve on the Kaplan Meier graph represents a different group of patients. Each group is called cohort. Patients are grouped together in two cohorts based on certain similarities which are shared by all patients. In this module patients are grouped together based on the type of cancer they have and also based upon the specific chemotherapy they have received in a clinical trial. The objective of comparing a chemotherapy treatment with chemotherapy is to see which treatments work better for patients who have the similar type of cancer.

Let's see how the Kaplan Meier graph works. The x-axis is the horizontal axis that representsthe time. In a cancer clinical trial time might be presented in weeks, months or years. The y-axis represents the numbers of patients. In the fig 1 one hundred patients receive treatment A.

In the beginning of the study all patients are alive. Each time a patient dies the line takes a tick downward to indicate that the number of patients still living has decreased. By the end of the study 75 people or 75 percent of people are still alive. In the second curve this group of patients are called cohort B and all patients in this cohort receive treatment B. Initially 100 patients receive treatment B and all people are alive in the start of this study. As the time moves out number of people alive diminishes. Finally, by the end about fifty patients or fifty percent of patients remain alive. Now, on comparing cohort A with cohort B a wide separation between the two curves is visible and at all points throughout the entire study more patients of cohort A are alive than cohort B. In simplest interpretation treatment A seems to be more effective than treatment B.

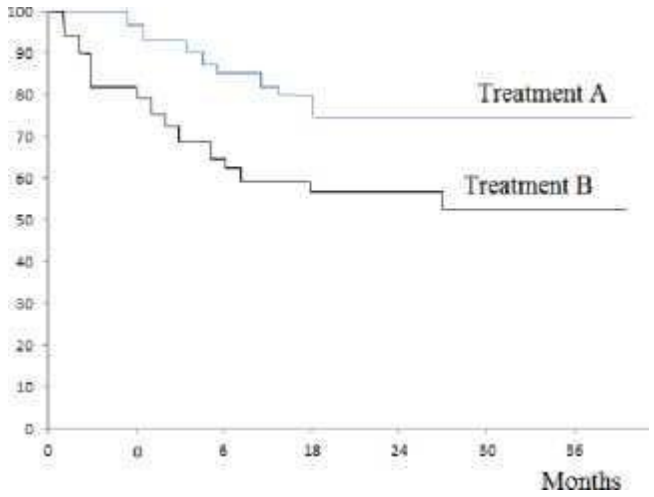


**Figure 1. Kaplan meier graph**

$$L2\ (logrank)= \frac{(A1\ B1\ 1}{g} + \frac{(Ai\text{-}Bi^2)}{B_2} \quad \textbf{(3)}$$

In above equation A1 and A2 are the aggregate count of observed events. B1 and B2 are the aggregate count of estimated events [5].

The total number of expected events during the occurrence of any event is called total number of expected events in the group [5]. The total number of expected events during the occurrence of any event is computed by calculating the product between chances of facing death and number of people survived in the group [5]. In order to calculate the test statistic author of paper calculated the sum of expected deaths in both the treatment groups 1 and 2 as well as observed number of deaths in group2. Finally, P value was calculated that differentiated the population survival curves very well. Therefore, log- rank test is popular method for comparing the difference in the survival curves which takes the whole follow-up period into consideration.

**3.2.4. Cox regression model:** Regression is a technique which is used to ascertain the intensive bond between the dependent and the independent variable [6]. When more than one independent variable need to be taken into the consideration, then this technique is known as multiple regression. Unlike multiple regression analysis, in Cox regression analysis the output variable is the hazard function at any given time. The sole objective of the model is targeted

 

to examine the various effects of several variables and determination of their endurance with time. A Cox's method helps in analyzing the effect of various treatments on the survival after adjustments of independent variables. Moreover, it is also used to examine the risk of death of a person on the basis of prognostic variables. The final model from Cox will produce a resultant equation which is a function of various independent variables called hazard function. The higher the regression coefficient, the higher are the chances of hazard which in turn decreases the chances of prognosis. Hazard function is the probability that a person will confront an event within an interval ensuring that person has lived till the starting of the interval. Hazard function is represented as h (t):

, , >. coun t p e o p 1 e c o n fro nti n g a n e v e n t i n i n t e rv a 1 s t a rt i n g a t t **(4)**
v J (count people living at time t)1(interval width)

## 4. Conclusions

Survival analysis is the study on survival of people suffering from various diseases over a period of time. In this paper various statistical methods have explained which are used to study right censored data. A Kaplan-Meier graph is extensively used to estimate and graph the survival probability of individuals as a function of time. Generally we use log rank test to examine the comprehensive differences in the survival curves of two or more groups like treated versus untreated groups but it is not allowed to examine the effect of other predictor variables on the survival time. However, it is not possible to measure the consequences of other independent variables on existing survival time. Therefore, Cox proportional model is used to examine the effect of other predictor variables on the survival time of the various groups of patients.

## References

[1] Seidel, Chris, "Introduction to DNA Microarrays", WILEY-VCH Verlag GmbH & Co. **(2008)**

[2] Eisen, B., Michael, Brown, O., Patrick,"DNA arrays for analysis of gene expression", Elsevier Inc, Vol. 303, pp. 179-205, **(1999)**

[3] Bewick, V., Cheek L., Ball, J., "Statistical review 12: survival analysis", Critical care, Vol. 8, No. 5, **(2004)**

[4] Mukaka, M.M, "Statistics Corner: A guide to appropriate use of Correlation coefficient in medical research", Malwai Medical Journal, Vol. 24, **(2012)**

[5] Bland, M., J, Altman, G, Douglas, "Statistics Notes The log rank test", BMJ Publishing Group Ltd, Vol. 328, **(2004)**

[6] Walters, J, Stephan, "What is a Cox model?" Hayward Medical Communications, **(2009)**

[7] D., Simona, "What is survival analysis?" Cornell University.