FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 5 No. 2, June, 2021, pp 327-333
DOI: https://doi.org/10.33003/fjs-2021-0502-553

INVESTIGATING THE RELATIONSHIPS BETWEEN EXPRESSED CANCER RELATED GENES AND SURVIVAL OF PATIENTS WITH BREAST CANCER

Chinenye N. and Bashir S.

Mathematical Science, Faculty of Physical Science, Federal University Dutsin-Ma
nchinenye@fudutsinma.edu.ng, bsule@gmail.com

ABSTRACT

Cancer stem cells are regulated by complex interactions with the components of the tumor microenvironment through networks of Cytokins and growth factors. These interactions are mediated by group of proteins and microRNAs (miRs), which are expressed or repressed. These expression levels are critical for cancer stem cell formation and expansion, enabling the promotion of tumor cell proliferation and migration, as well as for the survival of cancer recurrence and patient survival. Micro array and RNA deep sequencing (RNA-seq) provide tools with ability to generate transcriptome information, deciphering global gene expression patterns and quantifying a large dynamic range of expression levels. In this study 94 breast cancer patients were investigated based on miR and mRNA expression levels in which WDR1, APC and AKAP13 genes were identified as genes that play important role in the survival of patients and these genes differed significantly with respect to survival of patients. We used the Pearson correlation to identify the over-expressed and under-expressed genes. We demonstrated that parametric survival models can be used to model outcomes for breast cancer, and for our dataset the log-normal model demonstrated the best fit compared to other parametric models. Through the use of log-normal model, we examined how each of the identified genes influence the survival of breast cancer patients.

Key words: Cancer stem cells, microRNAs, Micro array and RNA deep sequencing, Pearson correlation

INTRODUCTION

Cancer is a disease which occurs when changes in a group of normal cells within the body leads to uncontrolled growth causing lump called a tumour. If these tumours are not treated, they can grow and spread into the surrounding tissue, or to the parts of the body through the lymphatic system and bloodstream, and can cause harm to the digestive, nervous and circulatory system Parkin D.M, Bray F., Ferlay J., and Jemal A.. (2012). There are different types of cells in the body that perform different functions, but they are essentially similar. They all have a control called a nucleus. The nucleus are chromosomes made up of what is known as deoxyribonucleic acid (DNA) Stewart, B. and Wild, C.P. (2014). DNA contains thousands of genes, which are coded messages that instruct the cell on how to function. Each gene is an instruction that tells the cell to make a protein, or a different type of molecule called ribonucleic acid (RNA). Together, proteins and RNA control the cell. They decide what sort of cell it will be, what it does, when it will divide, and when it will die. For a cancer to occur in the body, certain changes take place within the genes of a cell or a group of cells Stewart, B. and Wild, C.P. (2014). The change that occur is called mutation. It means that a gene has been damaged, lost or copied twice Some mutations means that the cell no longer understand its instructions and starts to grow out of control. There have to be about half a dozen different mutations before a normal cell turns into a cancer. Mutations in particular genes may mean that too many proteins are produced that trigger a cell to divide or proteins that normally tell a cell to stop dividing may not be produced. Breast cancer is a major public health problem across the world and it is the most common trespassing cancer in females in both developed and developing countries. It is the most frequently diagnosed cancer and the leading cause of cancer death among females worldwide, with an estimated 1.7 million cases and 521,900 deaths in 2012 Zhang Z., Yamashita H., Toyama T., Sugiura H., Omoto Y., Ando Y., Mita K., Hamaguchi M., Hayashi S., Iwase H. (2004). Breast cancer alone accounts for 25% of all cancer cases and 15% of all cancer deaths among females. Developed countries account for about one-half of all breast cancer cases and 38% of deaths. Rates are generally high in Northern America, Australia/New Zealand, and Northern and Western Europe; intermediate in Central and Eastern Europe, Latin America, and the Caribbean; and low in most parts of Africa and Asia. Torre L.A., Bray F., Siegel R.L., Ferlay J., Lortet-Tieulent J, and Jemal A. (2012). Hien-Wei T, Wen-Hung K, Shih-Hsuan C, Hong-Lin C, King-Jen C and Lu-Hai W (2018) used the transketolase (TKT) expression correlated with tumor size in the 4T1/BALB/c syngeneic model and discovered that TKT expression was higher in lymph node metastases compared with primary tumor or normal tissues of patients, and high TKT levels were associated with poor survival.

Ji-Eun K., Jaesung C., JooYong P., Chulbum P., Se Mi L., Seong E.P, Nan S., Seokang C., Hyuna S., Wonshik H., Jong W.L., Sue K. P., Mi Kyung K., Dong-Young N., Keun-Young Y., Daehee K. and Ji-Yeob C. (2018) found out that *ABCB1* rs1045642 was associated with poor progression-free survival in a meta-analysis (HR = 1.33, 95% CI: 1.07–1.64). *ABCB1*, *SLC8A1*, and *SLC12A8* were associated with breast cancer survival in SEBCS ($P_{\text{gene}} < 0.05$).

Gibbs, L.D., Chaudhary, P., Mansheim, K. *et al.* (2019) analyzed TCGA breast cancer database (*n* = 1098) to observe AnxA2 expression within breast cancer subtypes and is correlation with overall survival. They examined breast tissue specimens (*n* = 119) through chromogenic in situ hybridization (CISH) and specimen were scored independently by two pathologists in a blinded study and found out that high expression of AnxA2 was correlated with poor survival in patients with TNBC. AnxA2 gene expression was not correlated with poor survival in other breast cancer subtypes. AnxA2 average CISH intensity score (CISH score = 0, null expression to 3, high expression) for TNBC was significantly higher in comparison to estrogen receptor and/or progesterone receptor positive, human epidermal growth factor positive, and non-malignant tissues.

Huan H. (2019) investigated the genome-wide assessment of c-Jun, a potent member of AP-1 family, and ERα cistrome and transcriptome in ERα-positive breast cancer cells. Our findings demonstrate the genome-wide co-localization of ERα and c-Jun binding regions and suggest that ERα tethering to AP-1 is a global mechanism for gene transcription regulated by ERα. In addition, the results confirm that the sensitivity of ERα-positive breast cancer cells to tamoxifen therapy is reduced by c-Jun overexpression. He also showed that expression of transforming growth factor β induced (TGFBI) protein is associated with poor outcomes of ERα-positive breast cancer patients receiving endocrine therapy and thus as a candidate gene that may cause tamoxifen resistance through ERα and AP-1 crosstalk.

Bertucci F., Nasser V., Granjeaud S., Eisinger F., Adelaide J., Tagett R., Loriod B., Gi-aconia A., Benziane A., Mahdi, A. F., Malacrida, B., Nolan, J., McCumiskey, M. E., Merrigan, A. B., Lal, A., … Kiely, P. A. (2020) conducted a mass spectrometry screen following bioorthogonal noncanonical amino acid tagging to elucidate changes in the nascent proteome that occur during epidermal growth factor stimulation in migrating and invading cells.

Most of the previous work were motivated on the use of either mRNA or miR expression levels for breast cancer. However in this study we focused on the use of mRNA and miR expression levels of breast cancer patients. The univariate (exponential, Weibull, Log-logistic, log-normal and extreme-value distributions) were employed for the data set and expression level based model was developed to predict the survival of breast cancer patients.

Keywords: Breast cancer, gene expression and survival analysis.

Material and Methods

Breast Invasive Carcinoma (BRCA) dataset from The Cancer Genome Atlas (TCGA) (2012) was used for experiments including mRNA expression data, miR expression data and clinical data. The main clinical variable used are Patients barcode, days to death, cancer status, and vital status. Patients were considered eligible for the present study through the use of the expression levels of genes (miR and RNA expression levels, 94 patients were considered eligible out of 869 samples. Samples in TGCA are labelled with unique digit identifier without referring to the patients' name.

Survival analysis

Survival analysis is the modelling of lifetimes through the use of appropriate probability density function (p.d.f) which is denoted by f(t) and the commulative distribution function (c.d.f) denoted as F(t); which is the probability that the event has occurred over a given period of time say t. It helps in the visualisation of the time duration on the occurrence of a specific event such as time to death of patients with certain disease, remission duration of certain disease in clinical trial, incubation times of certain disease, failure time of certain manufactured products, and life times of elderly in particular social programs and so on. In survival analysis failure is used to define the occurrence of an event of interest. The distribution used for this work are exponential, Weibull, Log-normal, Log-Logistic and extreme value distributions.

Exponential distribution

The exponential distribution is characterised by one parameter λ known as the constant hazard rate. A high value of λ shows that there is a low risk and long survival time. The survival time follows an exponential distribution with a parameter λ. The probability density function is written as

$$f(t) = \lambda \exp(-\lambda t) \qquad 1.$$

where t ≥ 0, λ > 0. Given the probability density function above, from Equation (1) we have that cumulative distribution function is written as

$$F(t) = \int_0^t \lambda \exp(-\lambda x)dx = 1 - \exp(-\lambda t). \qquad 2.$$

It follows that the survival function is derived as

$$S(t) = 1 - [1 - \exp(-\lambda t)] = \exp(-\lambda t) \qquad 3.$$

and the hazard function

$$h(t) = \lambda \exp(-\lambda t) \exp(-\lambda t) = \lambda. \qquad 4.$$

Weibull distribution

The Weibull distribution is the general form of exponential distribution. It is characterised by γ and λ. The values of γ and λ determines the shape and scaling of the distribution curve respectively. The probability density function for a Weibull distribution is written as

$$f(t) = \lambda\gamma(\lambda t)\gamma - 1 \exp(-\lambda t)\gamma \qquad 5.$$

where t ≥ 0, γ > 0 and λ > 0. Given the probability density function, we derive the cumulative distribution function as

$$F(t) = \int_0^t \lambda\gamma(\lambda x)\gamma - 1\, exp(-\lambda x)\gamma dx = 1 - exp(-\lambda t)\gamma \qquad 6.$$

It follows that the survival function is derived as

$$S(t) = 1 - (1 - \exp(-\lambda t)\gamma) = \exp(-\lambda t)\gamma. \qquad 7.$$

and the hazard function

$$h(t) = \lambda\gamma(\lambda t)\gamma - 1 \exp(-\lambda t)\gamma \exp(-\lambda t)\gamma = \lambda\gamma(\lambda t)\gamma - 1. \qquad 8.$$

Log-normal distribution

The log-normal distribution is the distribution of a variable whose logarithm follows a normal distribution. The survival time T is said to be log-normally distributed if logT is normally distributed.

$$f(t) = \frac{\lambda\exp(-\gamma^2(log(\lambda t))^2)}{t\sqrt{2\pi}} \qquad 9.$$

where t ≥ 0, γ, λ > 0. Given the probability density function, the cumulative distribution function is derived as

$$F(t) = \int_0^t \frac{\lambda exp(-\alpha^2(log(\lambda x))^2)}{t\sqrt{2\pi}} \qquad 10.$$

It follows that the survival function is derived as

$$S(t) = 1 - \Phi(\gamma \log(\lambda t)).$$

Where $\Phi(t) = \frac{2}{\pi}\int_0^t exp(-x^2)dx$. And the hazard function is

$$h(t) = \lambda\exp(-\gamma 2(\log(\lambda t))2)\; t\sqrt{2\pi} = \Phi(\gamma\; log(\lambda t))$$

Log-Logistic distribution

The survival time T has a log-logistic distribution if logT has a logistic distribution. The log-logistic distribution is characterised by two parameters, λ and γ. The median of the log-logistic distribution is $\lambda^{-\frac{1}{\gamma}}$. It's probability density function is as follows

$$f(t) = \frac{\lambda\gamma t^{\gamma-1}}{(1+\lambda t^\gamma)^2}$$

The cumulative distribution function is written as

$$F(t) = \int_0^t \frac{\lambda\gamma t^{\gamma-1}}{(1+\lambda t^\gamma)^2} dx \qquad 11.$$

It follows that the survival function is derived as

$$S(t) = 1 - \left(\frac{\lambda t^\gamma}{1+\lambda t^\gamma}\right) = \frac{1}{1+\lambda t^\gamma}. \qquad 12.$$

And the hazard function

$$h(t) = \frac{\lambda\gamma t^{\gamma-1}}{(1+\lambda t^\gamma)^2}(1+\lambda t^\gamma) = \frac{\lambda\gamma t^{\gamma-1}}{1+\lambda t^\gamma} \qquad 13.$$

Extreme value distribution

The extreme value distribution arise as limiting distributions for maximums or minimums (extreme values) of a sample of independent, identically distributed random variables, as the sample size increases. The probability density function for extreme value distribution is written as

$$f(t) = \exp\left(\frac{t-\mu}{\beta}\right) \exp\left(-\exp\left(\frac{t-\mu}{\beta}\right)\right) \qquad 14.$$

where μ and β are constants. Using the cumulative distribution function we can get the probability density function as

$$F(t) = 1 - \exp\left(-\exp\left(\frac{t-\mu}{\beta}\right)\right) \qquad 15.$$

It follows that the survival function is derived as

$$S(t) = \exp\left(-\exp\left(\frac{t-\mu}{\beta}\right)\right) \qquad 16.$$

And the hazard function

$$h(t) = \exp\left(\frac{t-\mu}{\beta}\right) \qquad 17.$$

Estimation of parameter

The likelihood function of a specific parameter say $\beta = \beta_1, \beta_2 ... \beta_J, j$ unknown parameters is the joint probability density (or mass) function of a specific set of survival times say t_i, we denote the likelihood function as l(β) written as

$$l(\beta) = \log \sum_{1}^{n} f(t_i | \beta) \qquad 18.$$

where n is the number of points to be observed. We need to find the value of the constant β that maximizes this function, and that value is known as the maximum likelihood estimate.

Kaplan Meier Model. For a dataset with observed failure times$t_i = t_1, ..., t_r$ where r is the number of distinct failure times observed in the data, the Kaplan Meier estimate at any time is given by

$$S(t) = \prod_{j|t_i \leq t} \left(\frac{n_j - d_i}{n_j}\right) = \prod_{j=1} \left(\frac{n_j - d_i}{n_j}\right) \qquad 19.$$

where n_j is the number of individuals at risk at time t_i and d_i is the number of failures at time t_i. The product is over all observed failure times less than or equal to the time t.

RESULTS AND DISCUSSION

Making use of the samples of people with tumour, we compared the performance of the different models. The models were compared by the use of the Akaike Information Criterion (AIC) in other to select the best model that fits the data set. The model with the lowest AIC value corresponds to the best fit model. The result including the Chi-square, p-value, scale, log-likelihood of the model and the log-likelihood of the intercept is given in Table 1.

Distribution	Chi-Square	P-value	Log like(Model)	Scale	Log like(intercept)	AIC
Exponential	78.72	$5.1 \times 10^{-0.5}$	−448.7	1	-488.1	934.4267
Weibull	84.46	$9.1 \times 10^{-0.6}$	−444.9	0.699	-487.1	965.8274
Log-Logistic	77.69	$6.8 \times 10^{-0.5}$	−445.7	0.543	-484.5	967.3649
Log-normal	77.07	$8.2 \times 10^{-0.5}$	−444.1	0.915	-482.7	964.2974
Extreme-value	86.81	$4.4 \times 10^{-0.6}$	−478.4	1366	-521.8	1032.862

Table 1: Summary for the exponential, weibull, lognormal, log-logistic and extreme value distributions with n = 94: The table shows the Chi-square p-value, Log-like (Model), Scale, Log-like (intercept) and AIC for the distributions.

The model with the lowest AIC value is the exponential model with 934.4267 AIC value as seen in Table (1), but we cannot use the exponential model because the hazard function is constant which means that it does not depend on time and this is not appropriate for the cancer study. So, we pick the next model which is the Log-normal model. Because the log-normal model demonstrated a good fit for our data set, we select the model and then obtain the regression results for the model to see the genes that contribute to patient's survival. Using the Kaplan Meier Model the genes identified are WDR1, APC and AKAP13 the results obtained are given Figure 1, 2 and 3.

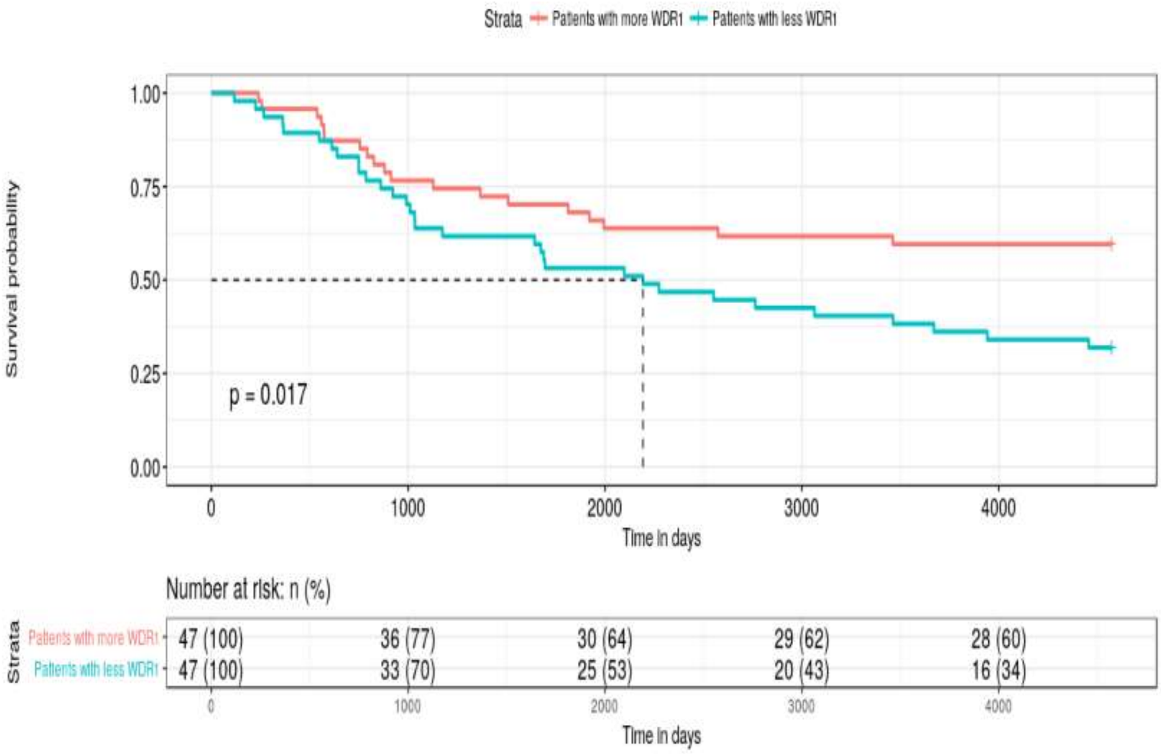

Figure 1 shows the survival in patients with WDR1: Kaplan Meier survival estimates in the 47 patients with WDR1.

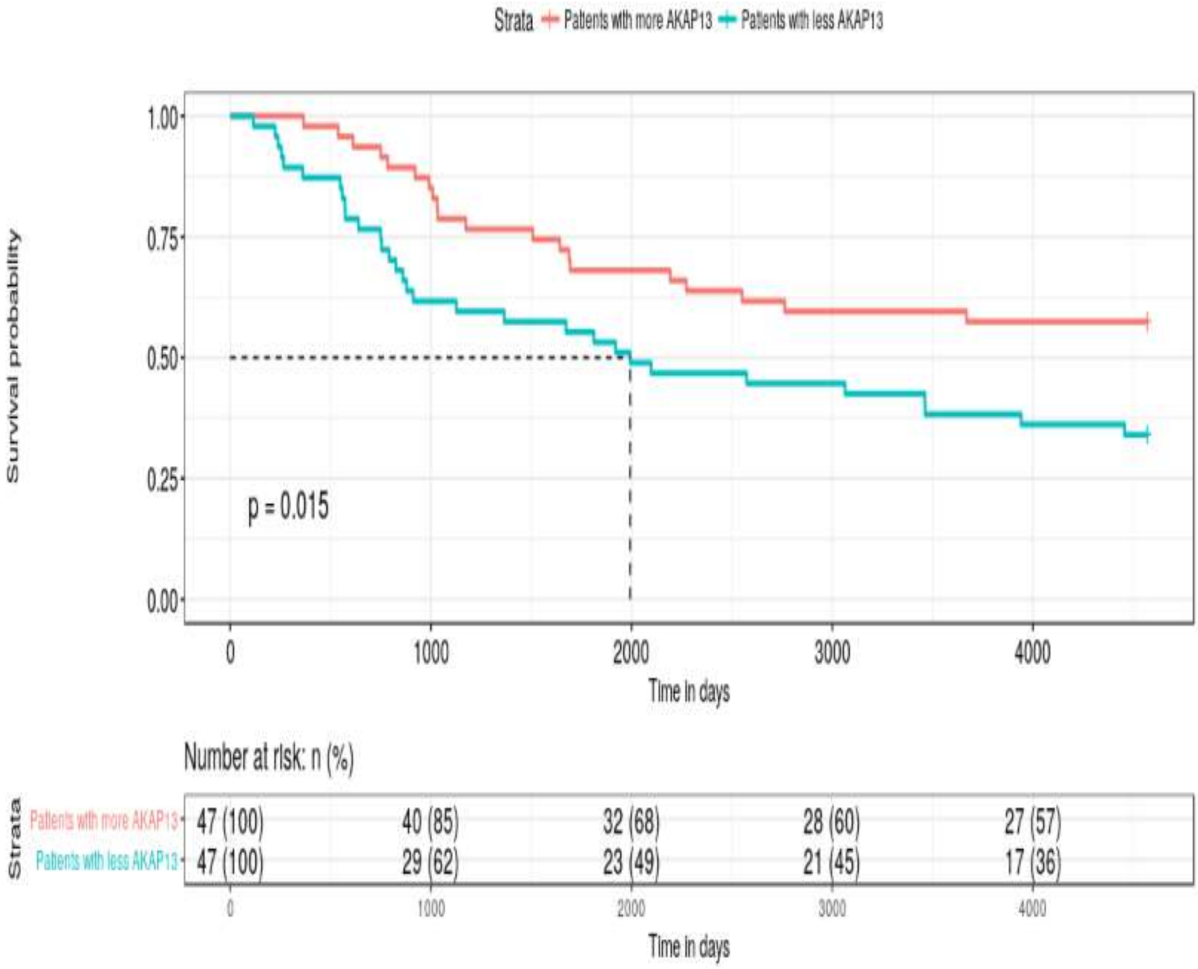

Figure 2 shows the survival in patients with AKAP13: Kaplan Meier survival estimates in the 47 patients with WDR1.

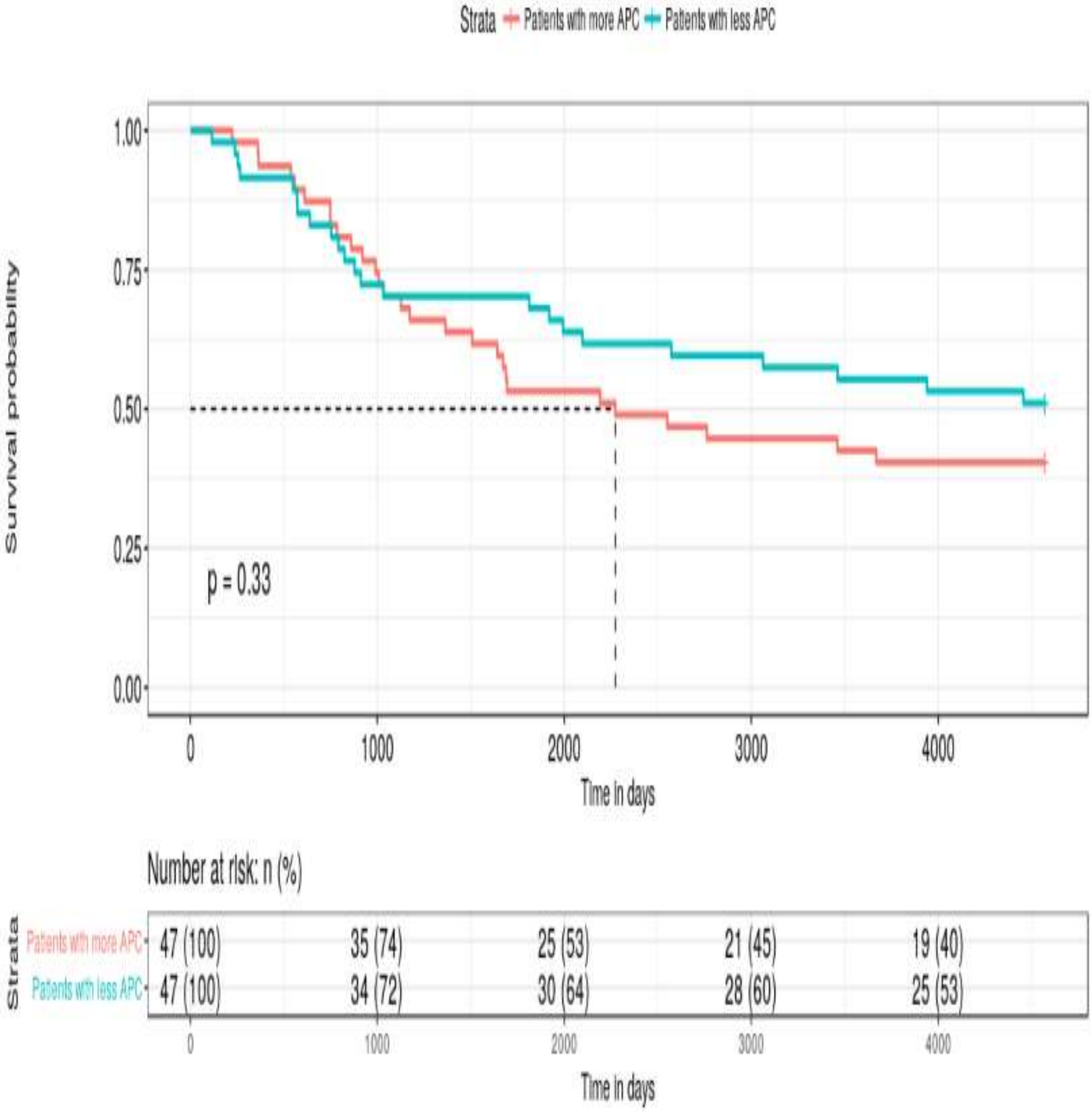

Figure 3 shows survival in patients with APC: Kaplan Meier survival estimates in the 47 patients with APC.

The Kaplan-Meier plot in Figure (1) shows that the probability of patients with high WDR1 expression level surviving is higher than those with low WDR1 expression level. The p-value, $p = 0.017 < 0.05$, shows that there is significant difference between patients with high WDR1 expression level and those with low WDR1 expression level. Also, the Kaplan-Meier plot in Figure (2) shows that the probability of patients with high AKAP13 expression level surviving is higher than those with low AKAP13 expression level. The p-value, $p = 0.015 < 0.05$, shows that there is significant difference between the patients with high AKAP13 expression level and that of those with low AKAP13 expression level. Moreover, the Kaplan-Meier plot in Figure (3) shows that the probability of patients with less APC expression level survival is higher than those with high APC expression level. The p-value, $p = 0.33 > 0.05$, shows that there is no significant difference between the patients with high APC expression level and that of those with low APC expression level.

Conclusion

From our result it is clear that AKAP13, WDR1 and APC are the genes that affects the survival of cancer patients'. Therefore, the understanding of the results can help in choosing or administering treatment methods aimed at increasing the expression levels of the AKAP13, WDR1 genes and decreasing of APC expression level of the cancer patients to help extend their survival period. It is also very important for the development of effective therapeutic strategies to tackle cancerous cells in the body system.

REFERENCES

Ferlay J., Soerjomataram I., Dikshit R., Eser S., Mathers C., Rebelo M., Parkin D.M, Forman D., and Bray F. (2012). Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan. International journal of cancer, 136(5):E359-E386.

Gibbs, L.D., Chaudhary, P., Mansheim, K. *et al.* (2019). ANXA2 expression in African American triple-negative breast cancer patients. *Breast Cancer Res Treat* **174,** 113–120. https://doi.org/10.1007/s10549-018-5030-5.

Hien-Wei T, Wen-Hung K, Shih-Hsuan C, Hong-Lin C, King-Jen C and Lu-Hai W (2018). Transketolase Regulates the Metabolic Switch to Control Breast Cancer Cell Metastasis via the α-Ketoglutarate Signaling Pathway. 10.1158/0008-5472.CAN-17-2906.

Huan H. (2019). A multi-omics approach to uncover estrogen receptor (ER) and activator protein 1 (AP-1) signaling networks in breast cancer

Mahdi, A. F., Malacrida, B., Nolan, J., McCumiskey, M. E., Merrigan, A. B., Lal, A., Kiely, P. A. (2020). Expression of Annexin A2 Promotes Cancer Progression in Estrogen Receptor Negative Breast Cancers. *Cells*, *9*(7), 1582. doi:10.3390/cells9071582.

Parkin D. M, Bray F., Ferlay J., and Jemal A. (2012). Cancer in Africa. Cancer Epidemiology and Prevention Biomarkers, 23(6):953-966.

Parkin D. M, Bray F., Ferlay J. and Jemal A. (2012). Cancer Epidemiology and Prevention Biomarkers, 23(6):953-966.

Stewart, B. and Wild, C.P. (2014). A global view of cancer, including cancer patterns, causes, and prevention.

The Cancer Genome Atlas (TCGA) (2012). Comprehensive molecular portraits of human breast tumours.

Torre L.A., Bray F., Siegel R.L., Ferlay J., Lortet-Tieulent J, and Jemal A. (2012). Global cancer statistics. CA: a cancer journal for clinicians, 65(2):87-108.

Ji-Eun K., Jaesung C., JooYong P., Chulbum P., Se Mi L., Seong E.P, Nan S., Seokang C., Hyuna S., Wonshik H., Jong W.L., Sue K. P., Mi Kyung K., Dong-Young N., Keun-Young Y., Daehee K. and Ji-Yeob C. (2018). Associations between genetic polymorphisms of membrane transporter genes and prognosis after chemotherapy: meta-analysis and finding from Seoul Breast Cancer Study (SEBCS)

Zhang Z., Yamashita H, Toyama T., Sugiura H., Omoto Y., Ando Y., Mita K., Hamaguchi M., Hayashi S. and Iwase H. (2004). HDAC6 Expression is Correlated with Better Survival in Breast Cancer.