Impact of Missing Data in Survival Analysis

***PAVITHRA V and **KANNAN R**

Department Of Statistics, Annamalai University

Abstract:*In statistical analysis, the problem of missing data deserves special attention because it refers to the case where not all data were obtained as intended in the study design. In survival analysis researchers are faced with the problem of identifying subjects for each follow-up visit and in some situations; may not obtain observations about the subject of the study. As a result, there will be a lack of data in some studies and this poses a major challenge for the analysis. In general, missing data analysis deals with replacement and missing data with deletion. In this paper may tried missing at random with deletion and Imputation techniques and focuses on the study is to find out the survival probability, rate and final outcome of the Breast Cancer patients by the method of Kaplan-Meier, Cox Proportional Hazard Model, Minimum Survival probability, Maximum Survival Probability.*

Keywords: *Breast Cancer, Missing Data, Kaplan-Meier Test, Log Rank Test, Cox Proportional Hazard Model,Minimum Survival probability, Maximum Survival Probability.*

1. Introduction:

Breast cancer (BC) is the non-communicable diseases and that begins in the cells of the breast. BC is one of the leading cancer among Indian women, with over 1.5 lakh new BC patients recorded in India in 2018. It accounts for 14 percent of all cancers among women BC is not common in men, with 1 in 400 men getting BC. This is most common in Indian women 1 in 28 can develop BC at some point in their lives. Unfortunately, the number of BC cases reported each year is increasing faster than ever. The BC accounts for more than 27 percent of all new cancer cases. There is an increase in the trend of new cancer patients and comparably the risk is higher in urban areas as 1 in 22 women and lower in rural areas as 1 in 60 women. In India, the average age of the high-risk group is 40-55 years are more prone to BC. The overall numbers in India are better compared to the number for developed countries like US/UK is less where in 1 in 8 women are diagnosed annually. However, as the awareness level about the disease in developed countries is quite high and there is a lot of government funding which promotes timely detection, most cases are detected and treated at early stages leading to better survival rates.

In India on the other hand, has a very low survival rate due to its large population and low awareness. 1 in 2 women diagnosed with BC will die within the next five year. One of the main reasons for high mortality rates is lack of awareness, late diagnosis and absence of proper BC screening programme. Most of the BCs are diagnosed at advanced stage. Many patients in the urban area are diagnosed at stage-2 and most of the cases from rural areas, these lesions are diagnosed only after they transform to metastatic tumors.The exact cause of BC is still unknown, but years of medical research have identified several risk factors. It is still unclear why some women at very high risk do not develop BC, while some women without risk factors may develop BC. The risk factors for BC include genetics and heritage, late pregnancy, use of oral contraceptives, early onset of menstruation, late menopause, excessive alcohol intake, smoking, adolescent obesity, increases stress and poor eating habits-these factors are due to the increased incidence of BC.

Through cancer, especially BC is a very dangerous disease that is widespread all over the world (Torre et al., 2015). Cancer is a group of diseases that causes by the uncontrolled growth and spread of abnormal cells anywhere in the body (Diabate et al., 2018). BC is so expensive that it has received a lot of consideration from doctors and statisticians. Mortality with unstable mortality with many different prognostic (Pg) factors (Parkin et al., 2014). American Joint Committee on Cancer (AJCC) BC staging is associated with survival prognosis (American Cancer Society., 2017). This situation is indicated by reduced survival from stage-1 90%, stage-2 65%, stage-3 20% and stage-4 5% (Sinaga et al., 2017).

The majority of BC cases are classified as either invasive or non-invasive. Invasive BC has spread throughout the body, but non-invasive not spread to the body (Abay et al., 2018). Age has a significant effect on women getting BC. The mortality rate from BC increases with increasing age (Rezaianzadeh et al., 2009). Drinking alcohol increases the risk of dying from BC in women is about 7% to 12% for every 10g of alcohol consumed per day (Desantis et al., 2013). A study conducted by Addis Ababa University on the impact of several risk factors on BC and survival showed that stage and disease type had a significant effect on survival of BC (Kantelhardt et al., 2014).

The problem of missing data was dealt with mainly by editing until 1970. Contribution to inference problems in missing data studies was developed by Rubin (1976). During the 1980's, several methods such as Listwise deletion, pairwise deletion, imputation methods and various models were widely applied. The missing data problem is very difficult to apply for the common statistical methods (Rubin, 1987). The studies of Little and Rubin (1987) mark the beginning of the second phase and can be considered a breakthrough in the development of missing data methods. This method is said to be better than the simple imputation method that provides efficient parameter estimates.

In some cases, missing data can lead to bias and lead to erroneous conclusions about changes in mean responses. Missing data will reduce the efficiency or accuracy of estimates of changes in mean. Unfortunately, the larger the amount of missing data, the greater the loss of model accuracy (Fitzmaurice et al.,(2004); Mohanraj and Srinivasan (2018)). Several reasons have been attributed to missing data, including equipment failure, audience attrition from treatment design or intrusive questions about a survey or unclear instructions.Fayers and Machin (2007) described a special case of single imputation called hierarchical scale imputation. Hedeker and Gibbons (2006), Fitzmaurice (2003) and Diggle et al., (2002) have made a valuable contribution to the development of the missing data problem. Hesketh and Skrondal (2008) provided a more applicable framework for STATA users to analyze missing data.The analysis of missing data poses a major problem because the estimates of the parameters are mostly biased (Becker and walstad (1990); Becker and Powers (2001); Holt (1997); Rubin (1976)). Studies by several researcher (Anderson et al., (1983); Kim and Curry (1977)) indicate a loss of information and statistical power when conducting analysis of missing data. Scheffer (2002) discussed how the mean and standard deviation are affected by different methods of imputation in dealing with different missingness mechanism.

1.1.Types Of Missing Data:

1) ***Missing At Random (MAR):*** The fact that the missing data is systematically linked to the observed data but not to the unobserved data.
2) ***Missing Complete At Random (MCAR):***The missing data is independent of observed and unobserved data. In other words, there was no systematic difference between participants with missing data and those with complete.
3) ***Missing Not At Random (MNAR):*** The fact that missing data is systematically linked to unobserved data, ie., the lack is related to events or factors not measured by the researcher.

In this article, we will calculate the survival probability and analysis by applying the MAR approach and using the three methods given below.

1.2.Methods Of Handling in Missing Data:

Ingeneral, the missing data analyzeunder three different techniques namely,

a) ***Pairwise Deletion:*** when the statistical procedure uses instances that contain missing data. A procedure cannot include a particular variable when it has a missing value, but it can still use the case when parsing other variables with a different value.
b) ***Listwise Deletion:*** in this method, an entire record is excluded from analysis if any single value is missing and therefore we have the same N (number of records) for all the analysis.
c) ***Single (or) Multiple Imputation:***in this manner,replace the missing value with single or multiple value using a strategy such as: Mean, Median, Most frequent,…,etc.

The goal of this study is to look into the survival and risk of death from the Adyar Cancer institute in 2013. We study the missing data in the received data from our point of view. Generally, the missing data divided into three types, namely, MACR, MAR and MNAR. Due to the presence of such missing data, the accuracy of the model is greatly reduced during the analysis and the researchers are confused when describing the model. Therefore, researchers have already used three methods namely, pairwise deletion, listwise deletion and imputation method to increase model accuracy when using such missing data. We are going to continue our survival analysis of BC research by using these three methods to overcome the deficiencies caused by missing data. This BC survival analysis implemented using, which includes various models, was employed. The Kaplan-Meier (K-M) with log rank test and Cox Proportion Hazard(PH) models are most commonly utilized models (Lee and Wang (2003)). In addition, we looked at the Minimum survival probability and Maximum survival probability (Felix and Kannan (2007)).

2.Statistical Methods:

2.1. Hazard Functions:

The hazard function of the hold time X is denoted by h(x) and defined as individual probability fails in the time interval $(x, x+\Delta x)$ that the individual has lived for time x, the hazard function is expressed as:

$$\boldsymbol{h(x)} = \lim_{\Delta x \to 0} \left[\frac{\boldsymbol{P(x < X < x + \Delta x | X > x)}}{\boldsymbol{\Delta x}} \right] \qquad \longrightarrow (1)$$

2.2. Cox Proportional Hazard Model:

The relationship between the hazard rate and the covariate set can be expressed using the model:

$$\ln[h(t)] = \ln[h_0(t)] + \sum_{i=1}^{n} x_i \beta_i \quad \longrightarrow (2)$$

Where $x_1, x_2, x_3, \ldots, x_n$ are covariates. $\beta_1, \beta_2, \beta_3, \ldots, \beta_n$ are the regression coefficients to be estimated. t is time and $h_0(t)$ is the baseline hazard rate when all covariates are zero.

2.3. The Survival Function:

Individual opportunities to survive for time x are expressed by S(x) = P(X > x). Let X be thecontinuous random variables, then the survival function is the complement of the CumulativeDistribution function S(x) =1-F(X) where F(X) = P(X ≤ x). The survival function is the integral of theprobability density function f(x):

$$\hat{S}(x) = P(X > x) = \int_{x}^{\infty} f(t)dt \quad \longrightarrow (3)$$

$$f(x) = -\frac{dS(x)}{dx} \quad \longrightarrow (4)$$

Then if X is the discrete random variables, and can be obtained $x_j$with the probability massfunction (p.m.f) $p(x_j) = P(X= x_j)$, j=1,2,3,… where$x_1, x_2, x_3, \ldots$then the survival function for thediscrete variables X is given by:

$$\hat{S}(x) = P(X > x) = \sum_{x_j > x} p(x_j) \quad \longrightarrow (5)$$

2.4. Kaplan-Meierwith Log Rank Test:

Estimated survival function for K-M Expressed as:

$$\hat{S}\left(x_{(j)}\right) = \hat{S}\left(x_{(j-i)}\right)\hat{P}\left(X > x_{(j)} \middle| X \geq x_j\right) \quad \longrightarrow (6)$$

In general, log rank is used to compare k-M survival curves formed by the following hypothesis:

H_0: There is no difference between the survival curves:

H_1: At least one difference between the survival curves:

$$\text{Log Rank Test} = \frac{(O_i - E_i)^2}{Var(O_i - E_i)} \quad \longrightarrow (7)$$

$$O_i - E_i = \sum_{j=1}^{n} m_{ij} - e_{ij} \quad \longrightarrow (8)$$

m_{ij} denotes the number of individuals who experience the event at timex_j, and e_{ij} is the value of hope.The null hypothesis will be rejected if log rank statistics≥ χ_α^2 with n-1 degrees of freedom (df) = 1 or p-value<α.

2.5.Minimum Survival Probability (MISP):

Survival probabilities are calculated on the assumption that all those that are censored, the result of interest occurred. Then, for any interval i, D_i denotes the number of

deaths during *i*, W_i denotes the number of censored observation during *i* and N_i denotes the number of subjects at the beginning of *i*. Then MISP for time interval *i* is expressed by

$$\text{MISP}=1-\frac{(D_i-W_i)}{N_i} \longrightarrow (9)$$

2.6. Maximum Survival Probability (MASP):

The survival probabilities are calculated by assuming that all those who are censored at time *i* are alive till the end of time interval *i*. Hence the notations of MASP is,

$$\text{MASP}=1-\left(\frac{D_i}{N_i}\right) \longrightarrow (10)$$

3. Source of Breast Cancer Data:

The data we used in our research were obtained from the Adyar Cancer Institute in Chennai. These data are the newly diagnosed breast cancer for 2013 and where we used the number 257 patients for our study. The data provided by the cancer center for this research: Gender, Age, Medical History, Date of Diagnosis, laterality of the BC, Grade, Stages, Treatments (Surgery, Chemo Therapy, Radiation Therapy, Hormonal Therapy) with dates, follow-up details with dates and Alive Status.

4. Result and Discussion:

4.1. Cox Proportional Hazard Model:

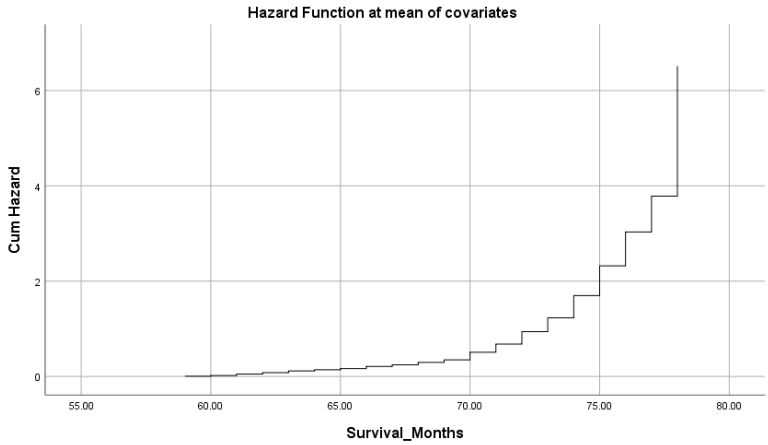

Figure-1: Hazard Function at mean of Covariate

The estimated variables by Cox regression are given in following: AgeGroup (X_1),RuralUrban (X_2),MedicalHistory (X_3), Laterality (X_4), Stage (X_5), Recurrence andMetastatic (X_6), Surgery (X_7), ChemoTherapy (X_8), RadioTherapy (X_9), HormonalTherapy (X_{10}).In step-1, the partial test shows that only Pg variables are statistically significant (P-value <5%). The backward stepwise method is used to extract the least influencing factors so that the final model is obtained in step-4 and the same method of Cox PH analysis applied all the three data set.

The β regression coefficient of the obtained models are all positive (β>0) with the value of exp(β)>0,meaning that all factors included in the model influence the event speed (death). That is, the risk of failure of depending on advanced stage of BC is 1.62 times

greater than those lower stages. The risk of death of BC patients with recurrent and metastatic is 0.623 times greater than those that do not have recurrent and metastatic.

Table-1: Partial Test with Backward Stepwise Method for deletion and imputation data set

	Indepented Variables	Pairwise Deletion				Listwise Deletion				Imputation Method			
		B	Wald	Sig.	exp(β)	β	Wald	Sig.	exp(β)	β	Wald	Sig.	exp(β)
Step-1	X_1	-.15	2.85	.092	.864	-.02	.053	.817	.978	-.146	2.850	.091	.864
	X_2	-.11	.43	.513	.895	.23	1.348	.246	1.253	-.118	.481	.488	.889
	X_3	-.06	.13	.718	.938	.01	.002	.961	1.010	-.055	.096	.757	.947
	X_4	-.07	.23	.633	.929	-.07	.162	.687	.930	-.070	.203	.652	.933
	X_5	.55	11.43	.001	1.709	.26	2.207	.137	1.296	.551	12.11	.001	1.73
	X_6	-.43	2.78	.095	.653	-.35	1.285	.257	.706	-.480	3.583	.058	.619
	X_7	.23	.29	.588	1.256	-.28	.333	.564	.751	.306	.529	.467	1.35
	X_8	-.14	.23	.635	.872	-.41	1.557	.212	.662	-.130	.205	.651	.878
	X_9	.24	1.43	.232	1.274	.23	1.095	.295	1.262	.245	1.457	.227	1.27
	X_{10}	-.05	.08	.784	.950	-.11	.253	.615	.896	-.048	.065	.798	.953
Step-2	X_1	-.11	1.94	.164	.897	-.03	.103	.748	.972	-.107	1.864	.172	.899
	X_3	-.07	.16	.691	.935	.07	.112	.738	1.067	-.064	.146	.703	.938
	X_5	.46	11.63	.001	1.59	.19	1.707	.191	1.216	.480	12.41	.000	1.62
	X_6	-.41	2.48	.115	.673	-.39	1.578	.209	.682	-.448	3.227	.072	.639
Step-3	X_1	-.09	1.84	.175	.910	-.04	.291	.589	.959	-.093	1.783	.182	.911
	X_5	.47	11.83	.001	1.59	.19	1.659	.198	1.212	.484	12.63	.000	1.62
	X_6	-.39	2.47	.116	.673	-.38	1.557	.212	.684	-.448	3.226	.072	.639
Step-4	X_5	.47	11.71	.001	1.59	.19	1.733	.188	1.217	.481	12.49	.000	1.62
	X_6	-.42	2.83	.093	.656	-.38	1.552	.213	.685	-.474	3.63	.057	.623

Table-2: Overall Score in Backward Stepwise Method for deletion and imputation data set

Handling Of Missing Data	Stepwise Method	-2 Log Likelihood	Overall (score)		
			Chi-square	df	Sig.
Pairwise Deletion	Step-1	1566.067	22.947	10	.018
	Step-2	1569.285	19.361	4	.001
	Step-3	1569.431	19.112	3	.000
	Step-4	1571.208	17.311	2	.000
Listwise Deletion	Step-1	1152.007	8.572	10	.573
	Step-2	1155.995	4.827	4	.306
	Step-3	1156.107	4.729	3	.193
	Step-4	1156.398	4.432	2	.109
Imputation Method	Step-1	1570.503	25.330	10	.005
	Step-2	1573.573	21.383	4	.000
	Step-3	1573.730	21.134	3	.000
	Step-4	1575.568	19.399	2	.000

In Table-1: The result of cox proportional hazard analysis showed that the most significant pg variable to the probability of death was the presence of advanced stage and tumor recurrence with metastatic. Table-2 indicated the overall score for the data set of pairwise deletion and imputation methods are more appropriate in missing data analysis. Meanwhile, the Listwise deletion leads reduced the efficiency and lower the precision of the model estimates.

4.2.Kaplan-Meier Analysis:

TheK-M estimated the probability of survival curve for missing data technique of pairwise deletion, listwise deletion and imputation method. The following figures are according to the two impact variables of stages and recurrence with metastasis.

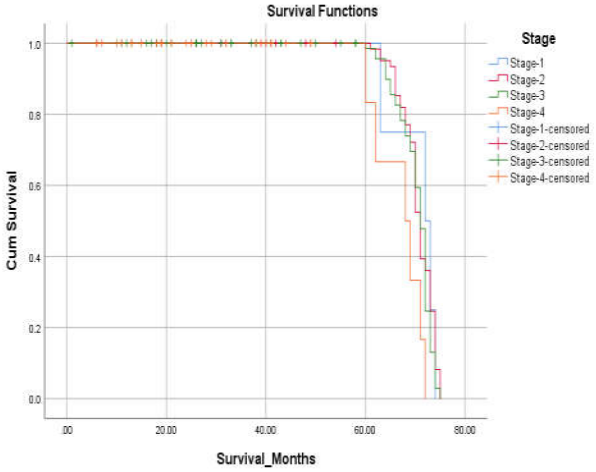

Figure-2: Tumor stage for Pairwise deletion

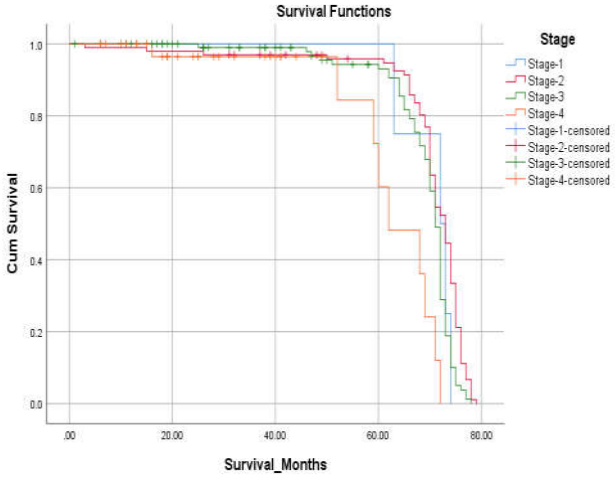

Figure-2: Tumor stage for Listwise Deletion

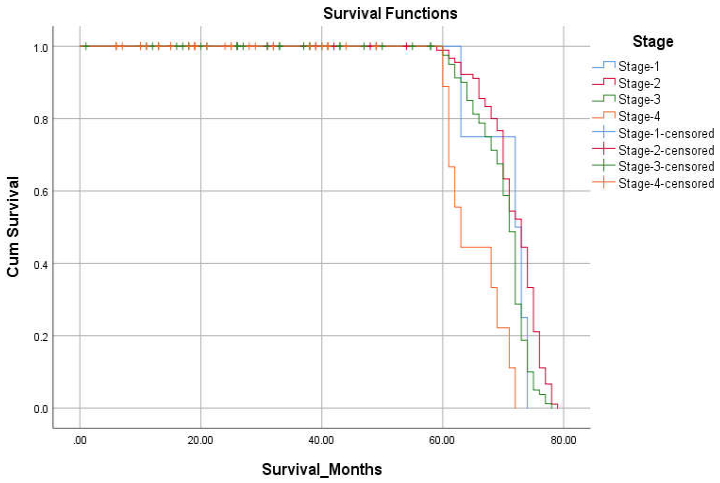

Figure-4:Tumor stage for Imputation methods

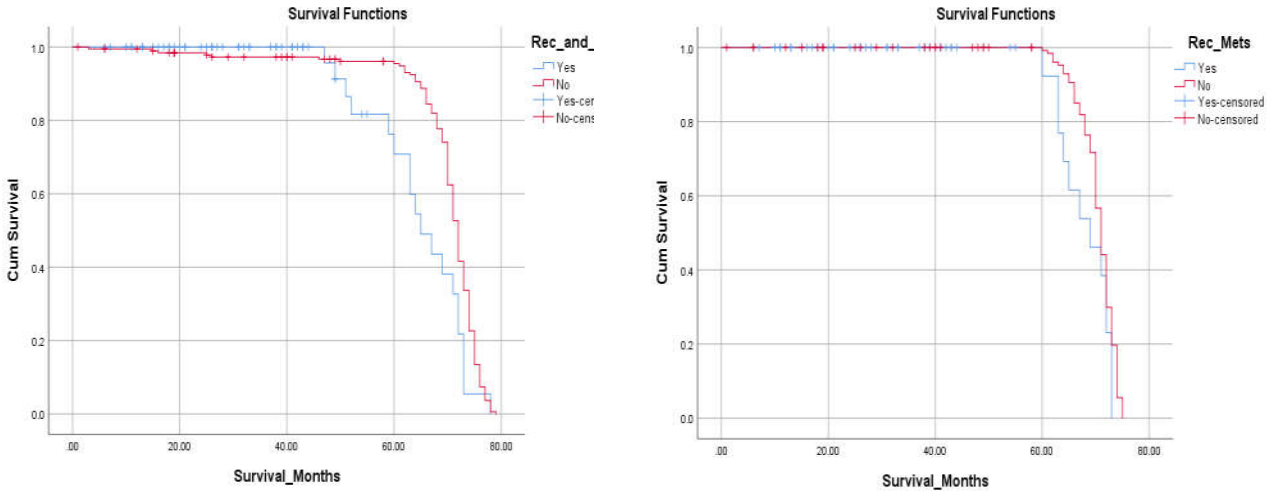

Figure-5&6: BCRecurrent and Metastaticpatientsfor Pairwise & Listwise Deletion

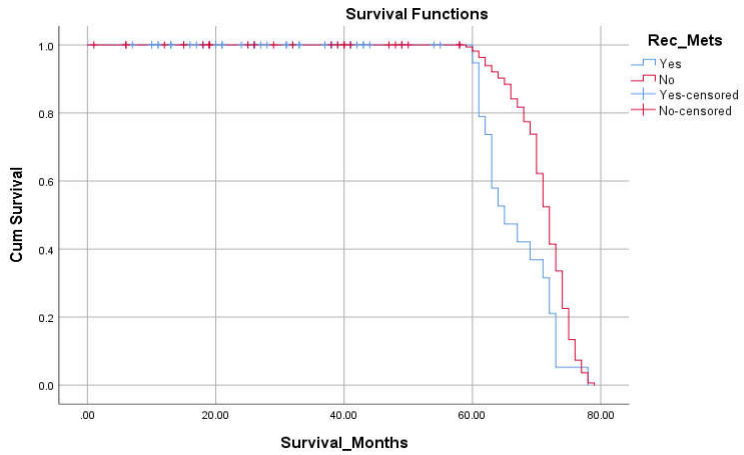

Figure-7:BC Recurrent and Metastatic patients for Imputation Method

The stages and recurrence with metastasis are the most important factor variable in theBC and these can determine the conditions of the cancer patients. In figure-2,3 &4 survival probability shown that the cancer Stages of patient with BC and clearly seen which stage is mostly survived or died in these BC. The survival rate of stage-1 and stage-2 patients was very high compared to stage-3 and stage-4 and risk rate of BC patients in stage-1 and stage-2 was very low when compared to other stages. In figure-5,6&7 survival probability shown that the recurrence with metastatic cancer patient and clearly seen who have not spread the cancer they are only mostly survived in these BC.

4.3.Log RankTest:

The log rank test to determine if there is a difference between the survivals curves. The log rank test of significant or not significant in Pg variables are given Table-2.

Table 3: Log Rank test used for Pgvariable affecting Survival of BC

Log Rank Test /Pg Variable	df	Pairwise Deletion		Listwise Deletion		Imputation	
		χ^2	Sig.	χ^2	Sig.	χ^2	Sig.
Age Group	4	4.518	.340	2.594	.628	4.493	.343
Area	1	.250	.617	.852	.356	.275	.600
Medical History	1	.067	.796	.319	.572	.064	.800
Laterality	1	.798	.372	.070	.791	.775	.379
Stage	3	25.353	.000	6.604	.086	29.523	.000
Recurrence & Metastatic	1	6.742	.009	3.342	.068	8.409	.004
Surgery	1	.713	.399	.587	.444	1.027	.311
Chemo Therapy	1	2.132	.144	3.750	.053	2.073	.150
Radiation Therapy	1	2.482	.115	.197	.657	2.643	.104
Hormonal Therapy	1	.137	.711	.133	.716	.149	.699

Based on the Log Rank Test in Table-3, the equality of survival distribution of the BC variables Cancer Stages and Recurrence with metastatic were statistically recorded a p-value (0.000 and 0.004) makes a significant difference and other variables have statistically no significant difference. Meanwhile, beauty of this study is in the handling the missing data technique, the imputation and pairwise deletion methods had outperforming when compared to listwise deletion method.

4.4.Comparison Of Survival probabilities:

Table-4. Survival Probability for Deletion and Imputation Methods

Handling Of Missing Data Technique	Method Of Probability	BC Five Year Survival Probability by Percentage				
		1st Year	2nd Year	3rd Year	4th Year	5th Year
Pairwise Deletion	MISP	95.3%	86.3%	79.7%	71.9%	66.5%
	MASP	95.7%	87.5%	81.7%	74.7%	71.2%
	K-M	95.4%	87.1%	80.3%	72.4%	69.3%
List wise Deletion	MISP	-	-	-	-	-
	MASP	94.8%	84.5%	77.5%	70.1%	65.4%
	K-M	93.4%	82.7%	75.8%	69.2%	64.4%
Imputation method	MISP	95.7%	87.5%	81.7%	74.7%	73.1%
	MASP	96.2%	88.3%	82.6%	75.3%	74.2%
	K-M	96.0%	87.9%	82.1%	74.9%	73.6%

Table-4: Shows the cumulative survival probabilities at the end of each year from the date of completion of treatment through different methods. These estimates are obtained by using MISP, MASP and K-M methods. In general, by all the methods estimates of the cumulative probabilities have been decreased as the survival period has increased. The higher probabilities have been estimated by MASP. i.e., the estimates of MISP and MASP provide the two extreme values of the survival band within which the true survival probability lies. The three estimates are similar but not identical. The overall five-year survival probability (%) for the BC patients has been found to be 74%, which is very much similar to other methods. However, this overall survival probability may not be an appropriate one, since the stage of the disease at diagnosis is one of the significant factors associated with the number of deaths occurred.

5. Conclusion:

The K-M, Cox PH, MISP and MASP survival results of the study showed that age, medical history, resident, laterality of breast, stage, recurrence, metastasis, surgery, chemo therapy, radiation therapy and hormone therapy affected the time to death of BC patients 2013 at Adyar Cancer Hospital. The K-M estimated the survival month of the BC is 70 months. The analyses Cox PH found main factor behind the poor survival time is that the treated patients is already in the advanced stage and recurrent with metastatic. The comparison between the MISP, MASP and K-M analysis the MASP and K-M showed similar together and most useful to survival analysis. The beauty of this survival analysis with missing data studies, here we have clearly outlined how to handling missing data in survival analysis. We reports that among the three methods used in this survival analysis, the pairwise deletion and imputation methods are more suitable for all type of analysis. The information loss is high and the model accuracy is very low for the listwise deletion method when compared to the other two methods. So it's best to avoid using the listwise deletion method when dealing with missing data.

6. Recommendation:

Health professionals, governments and NGO should raise awareness of early cancer screening and should also encourage women to be diagnosed at an early stage to improve mortality risk, and cancer screening facilitation and scheduling should be planned and scheduled in rural areas of this region to elucidated mortality risk.

Reference:

1. Abay M., Tuke G., Zewdie E., Abraha TH., Grum T and Brhane E. (2018). Breast self-examination practice and associated factors among women aged 20-70 years attending public health institutions of Adwa town, North Ethiopia. BMC Research Notes, Vol-11, Issue-1, 622, pp 1-7.

2. American Cancer Society. Stages of Breast Cancer (Internet-2017-2018).

3. Anderson AB., Basilevsky A., and Hum DPJ. (1983). Missing data: A review of the literature. In P.H. Rossi, J.D. Wright, & A.B. Anderson (Eds.), Handbook of survey research, (415-494). San Diego: Academic Press.

4. Desantis C., Siegel R., and Jemal A. (2013). Breast cancer facts and figures 2013-2014. American Cancer Society, pp 1-38.

5. Diabate M., Coquille L., and Samson A. (2018). Parameter estimation and treatment optimization in a stochastic model for immunotherapy of cancer. arXiv Preprint ArXiv, 1806.01915.

6. Becker WE and Powers JR(2001). Student performance, attrition, and class size given missing student data. Economics of EducationReview, Vol-20, pp 377-388.

7. Becker WE and Walstad WB (1990). Data loss from pretest to posttest as a sample selection problem. The Review of Economics and Statistics, Vol-72, Issue-1, pp 184-188.

8. Diggle PJ., Heagerty P., Liang KY., and Zeger SL. (2002). Analysis of longitudinal data (second edition). Oxford University Press. New York.

9. Fayers PM and Machin D (2007). Quality of life: the assessment, analysis and interpretation of patient-reported outcomes(Second Edition). West Sussex: John Wiley & Sons.

10. Felix AJ and Kannan R (2007). Statistical models in survival analysis, Chap-3, pp 50-51.

11. Fitzmaurice GM (2003). Methods for handling dropouts in longitudinal clinical trials. StatisticaNeerlandica, Vol-57, pp 75-99.

12. Fitzmaurice GM., Laird NM., and Ware JH. (2004). Applied Longitudinal Analysis. John Wiley and Sons. New York.

13. Hedeker D and Gibbons RD (2006). Longitudinal Data Analysis. John Wiley & sons.New Jersey.

14. Hesketh SR and Skrondal A (2008). Multilevel and Longitudinal Modeling Using Stata (2nd ed.). College Station, TX : Stata Press.

15. Holt D (1997). Missing data and nonresponse. In J.P. Keeve (Ed.) Educational research, methodology, and measurement: An international handbook (2ed.). New York: Elsevier Science Ltd.

16. Kantelhardt E., Zerche P., Mathewos A., Trocchi P., Addissie A., Aynalem A., Wondemagegnehu T., Ersumo T., Reeler A., Yonas B., Tinsae M., Gemechu T., Jemal A., Thomssen C., Stang A., & Bogale S. (2014). Breast cancer survival in Ethiopia: A cohort study of 1,070 women. International Journal of Cancer, Vol-135, Issue-3, pp 702-709.

17. Kim JO and Curry J (1977). The treatment of missing data in multivariate analysis. Sociological Methods & Research, Vol-6, Issue-2, pp 215-240.

18. Lee ET and Wang J (2003). Statistical methods for survival data analysis. Edition-3, John Wiley & Sons. New York.

19. Little RJA and Rubin DB (1987). Statistical Analysis with Missing Data. John Wiley& Sons, New York.

20. Mohanraj J and Srinivasan MR (2018). Missing longitudinal data analysis with covariance structure. Aligarh Journal Of Statistics, Vol-38, pp 83-102.

21. Parkin DM., Bray F., Ferlay J., & Jemal A. (2014). Cancer in Africa 2012. Cancer Epidemiology and Prevention Biomarkers, Vol-23, Issue-6, pp 953-966.

22. Rezaianzadeh A., Peacock J., Reidpath D., Talei A., Hosseini SV., and Mehrabani D. (2009). Survival analysis of 1148 women diagnosed with breast cancer in Southern Iran. BMC Cancer, Vol-9, Issue-1, 168.

23. Rubin DB (1976). Inference and missing data. Biometrika, Vol-63, pp 581-592.

24. Rubin DB (1987). Multiple Imputation for Non-response in Surveys. John Wiley & Sons.New York.

25. Scheffer J (2002). Dealing with missing data. Research Letters in the Information and Mathematical Sciences, Vol-3, pp 153-160.

26. Sinaga ES, Ahmad RA, and Hutajulu SH (2017). Berita kedokteranmasyarakat. Vol-33, FakultasKedokteran, Universitas Gadjah Mada.

27. Torre L. A., Bray F., Siegel RL., Ferlay J., Lortet-Tieulent J., & Jemal A. (2015). Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians, Vol-65, Issue-2, pp 87-108.