Research Paper
Impact Factor: 3.567
Peer Reviewed Journal

IJMDRR
E- ISSN –2395-1885
ISSN -2395-1877

CERVICAL CANCER ANALYSIS USING LOGISTIC REGRESSION IN THIRUCHIRAPALLI DISTRICT

S. Sasikala* **S. Vetrivel***
**PG & Research Department of Statistics, Periyar E.V.R College, Thiruchirapalli.*

Abstract
In this Study, an attempt is made to examine the influence of the factors age, family History, Education, occupation, tobacco uses, Diagnosis staging, Treatment therapy, Surgery, Follow up and the recovery of the patients from cervical cancer disease by using the statistical technique of Logistic regression. Data were obtained from september13 to august14 in Private Cancer hospitals in Thiruchirapalli district. The results were obtained by using logistic regression analysis. The significant risk factors for cervical cancer were identified. From the analysis, the diagnosis staging and tobacco uses turned out to be significant factors.

Key Words: Cervical Cancer, Prediction, Logistic Regression, Odd Ratio, Factors.

1. Introduction

Cervical cancer is a cancer arising from the cervix. It is due to the abnormal growth of cells that have the ability to invade or spread to other parts of the body. Early on, typically no symptoms are seen. Later symptoms may include abnormal vaginal bleeding, pelvic pain, or pain during sexual intercourse. While bleeding after sex may not be serious, it may also indicate the presence of cervical cancer. Human papilloma virus (HPV) infection appears to be involved in the development of more than 90% of cases; most people who have had HPV infections, however, do not develop cervical cancer. Other risk factors include smoking, a weak immune system, birth control pills, starting sex at a young age, and having many sexual partners, but these are less important. Cervical cancer typically develops from precancerous changes over 10 to 20 years. About 90% of cervical cancer cases are squamous cell carcinomas, 10% are adenocarcinoma, and a small number are other types. Diagnosis is typically by cervical screening followed by a biopsy. Medical imaging is then done to determine whether or not the cancer has spread. HPV vaccines protect against between two and seven high-risk strains of this family of viruses and may prevent up to 90% of cervical cancers. As a risk of cancer still exists, guidelines recommend continuing regular Pap smears. Other methods of prevention include: having few or no sexual partners and the use of condoms. Treatment of cervical cancer may consist of some combination of surgery, chemotherapy, and radiotherapy.

2. Materials and Methods

Sample and Procedure

The data for this study was collected from September13 to August 2014 by a concern department. Patients who had performed cancer cervical screening were selected about their history. Information from 150 patients was recorded based on their history and follow-up and results in the concern nominal register. Here ten variables are considered for classifying the presence of cervical cancer. The independent variables are patient age, family History, Education, occupation, tobacco uses, Diagnosis staging, Treatment therapy, Treatment Surgery, Follow-up. And Recovery is chosen to be the dependent Variable. Data analysis is performed using SPSS.20 Software.

3. Measures

All the data’s are coded.

4. Data Analysis

We choose a statistical technique, which is more apt to analyze the numerical evidence of the variables. Multiple Regression analysis and discrimination analysis are two related techniques pose difficulties to kind. These techniques pose difficulties when the dependent variable has only 2 values 0 and 1 the assumption are violated. For example, the variable is not reasonable to assume that the distribution of error is normal. Another bottleneck with multiple regression analysis is that predicted values cannot be interpreted as probabilities. Linear discriminate analysis does not allow direct prediction of group membership, but the assumption of multivariate normality of the independent variables and equal covariance matrices in the 2 groups are required for the prediction role to be optimal.

4.1 Logistic Distribution

The Logistic distribution is defined as

$$F_x(X) = [1 + \exp\{-X - a)/ \quad \}]^{-1}$$

The probability density function is also defined as

$$P_x(X) = \quad ^{-1}[\exp\{-(x-\sim)/\quad\}][1+\exp\{\text{-}(x\text{ - }\sim)/\quad\}]^{-2}$$

Rate of growth = [Excess over initial (asymptotic) value A] * [deficiency compared with final (asymptotic) value B) The solution of this equation is,

$$F(x) = \frac{Bde^{x/c} + A}{De^{x/c} + 1}$$

Where
D is constant as x → ∞, F(x) → A;
As x → ∞, F(x) → B;

The logistic distribution has a similar shape that of the normal distribution. It makes profitable on such occasions to replace the normal by the logistic to simplify the analysis.

In normal distribution β2=3 and logistic distribution β2= 4.1. This difference makes longer tails of the logistic distribution.

If the logistic distribution in used instead of normal distribution, to represent the population tolerance distribution then the analysis is carried out in terms of logics instead of probes.

Figure: 4.1.1 Patient History Statistics

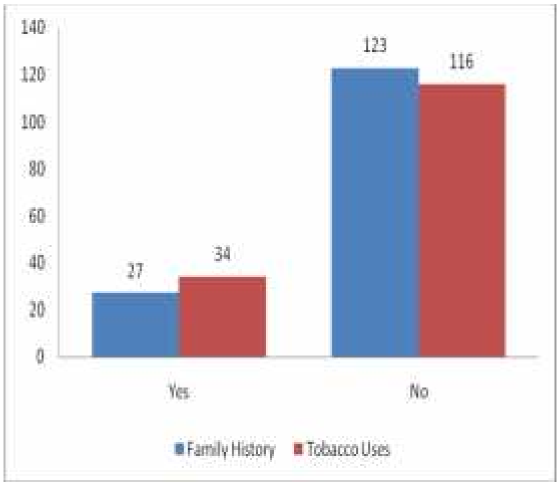

Figure: 4.1.2 Patient Diagnosis Stages

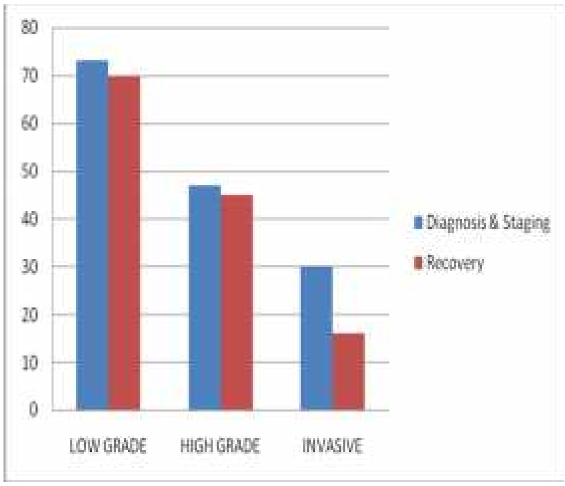

Table: 4.1.1 Percentage of Patient History

Factor	Yes	No
Family History	18%	82%
Tobacco Uses	22.7	73.3

Table: 4.1.2 Percentage of Diag. and Rec. rate

Factor	Low Grade	High Grade	Invasive Cacx
Diag. & Stag.	48.6%	31.3%	20%
Recovery	53.4%	34.3%	12.2%
Non Recovery	46.6%	65.7%	87.8%

Figure: 4.1.3 Patient Treatment and Recovery Statistics

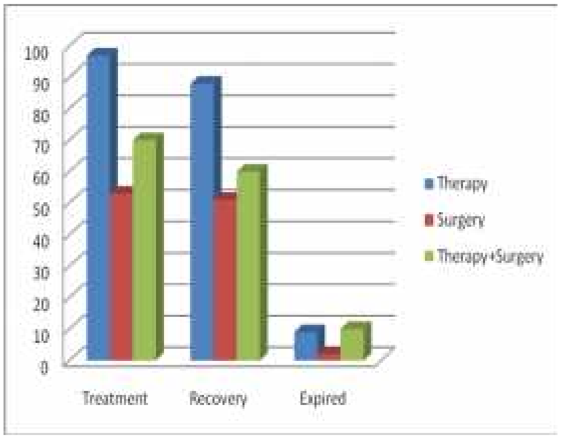

Fig: 4.1.4 Patient Diagnosis Stages and Recovery Rate

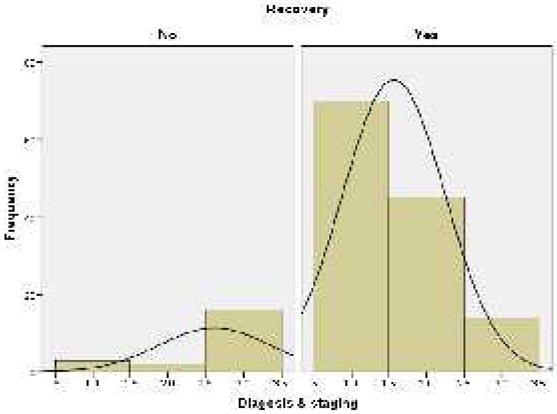

Table: 4.1.3 Patient Treatment and Recovery Statistics

Factor	Treatment	Recovery	Non Recovery
Therapy	64.6%	90.7%	9.3%
Surgery	35.3%	96.2%	3.8%
Surgery +Therapy	46.6%	85.7%	14.3%

4.2 Logistic Regression Model

In logistic regression model, the outcome (dependent variables) is binary or dichotomous. Logistic regression model is a multivariate technique for estimating the probability for the occurrence of an event.

For a single independent variable, the logistic regression can be written as probability that an event occurring is,

$$\text{Prob}_{\text{(event)}} = \frac{1}{1+e^{-z}}$$

Where
z = B_0+ B_1X and B_0 and B_1, are the estimated coefficience from the data, X is independent variable.
For more than one independent variable, the logistic model can be written as

$$\text{Prob}_{\text{(event)}} = \frac{1}{1+e^{-z}}$$

Where
$Z = B_0 + B_1X_1 + \ldots\ldots\ldots + B_nX_n$, is a linear combination of independent variables.

The probability of the event not occurring is estimated as

Prob (not Event) = 1 - Prob (Event)

If the estimated probability of the event is <0.5., we predict the event will not occur. If the probability is >0.5 we predict that the event will occur.

4.3 Odd of an Event

The odd of an event is defined as the probability of the outcome event occurring divided by the probability of the event not occurring.

The odds ratio for a predictor tells the relative amount by which the odds of the outcome change, when the value of the predictor value is increased by one -unit.

1. An odds ratio of the probability of the outcome being equal to 1.00 indicates that the independent variables.
2. An odds ratio larger (or) smaller than 1.00 indicates the recovery of patients from cancer being positively or negatively related to the independent variables that are being considered.

4.4 Finding

A multiple Logistic Regression analysis is carried out the result is given below.

Hosmer & Lemeshow Test

Hosmer and Lemeshow test is used to test the goodness of fit for the given model.

Hosmer and Lemeshow Test			
Step	**Chi-square**	**df**	**Sig.**
1	9.775	8	0.281

From the above table, the significances value is 0.281 which is greater than 0.05, the level of significances, the chosen logistic regression is a good fit for the given data.

4.5 Parameters Estimated

The parameters of the model are estimated by maximum likelihood methods. The following table 4.5.1 gives the variables in the equation.

Table: 4.5.1. Parameters Estimated Tables

Parameters	B	S.E.	Wald	Df	Sig.	Exp(B)
Age	.187	.651	.082	1	.774	1.205
Family History	1.232	.932	1.747	1	.186	3.428
Education	.421	.511	.679	1	.410	1.524
Occupation	-.603	.416	2.096	1	.148	.547
Tobacco Uses	1.684	.851	3.917	1	**.048**	5.385
Diagnosis Staging	-2.523	.762	10.973	1	**.001**	.080
Treatment_Therapy	-.689	1.324	.271	1	.603	.502
Treatment_Surgery	.220	1.204	.033	1	.855	1.246
Followup	-.756	1.279	.349	1	.554	.470
Constant	5.185	3.212	2.607	1	.106	178.638

From the above table, we observe that,

Z=**5.185**+0.187(Age) +1.232 (Family History) 0.421 (Education) -0.603 (occupation) +1.684 (Tobacco Uses) -2.523 (diagnosis) +-0.689 (Treatment therapy) 0.220(Treatment surgery) -0.756 (Follow-up).

5. Conclusion

From the analysis, we were obtained the following results,

1. The variables diagnosis and staging is found to be significant. The overall odd ratio for the variable diagnosis and staging is 0.080. Here, stage II and III odd ratio is 26.667, 25.714. It means the diagnosis and staging level increases the recovery rate is decrease.
2. The variables tobacco uses is found to be significant. The odd ratio for the variable tobacco uses is 5.385. It means tobacco uses increase means the recovery rate is decreases.
3. The variables age are found to be significant. It means the age level increases the recovery rate also decreases.
4. The variables education is found to be not significant. The odd ratio for the variable education is 1.524. It means the education level increases the recovery rate also increases.
5. The variables treatment (surgery) is found to be not significant. The odd ratio for the variable treatment (surgery) is 1.246. It means the treatment (surgery) level increases the recovery rate also increases.
6. The variables follow-up is found to be not significant. Its indicates treatment with follow-up monthly once review done means the odd ratio is 0.750. It means the follow-up increases the recovery rate also increases.

6. Limitation

As the study consists of 150 cases only, we could not generalize the effect. For large samples, the other variables age, family History, Education, occupation, tobacco uses, Diagnosis, Treatment suggested, Follow-up and the recovery also could have turned out to be significant.

7. Suggestions & Discussions

- Exposures to certain viruses, such as the human papilloma virus (HPV) and human immunodeficiency virus (HIV or AIDS), have been linked to an increased risk of developing certain types of cancers
- Cancer treatment is influenced by several factors, including the specific characteristics of your cancer; your overall condition; and whether the goal of treatment is to cure your cancer, keep your cancer from spreading, or to relieve the symptoms caused by cancer.
- Cancer disease is a major cause of death (and the number one cause of death in the world) of course one must also consider other factor such as lifestyle, for instance the amount of exercise one undertakes and their diet, as well as their overall health (mental and social as well as physical).
- It is controversial, but in my opinion that is because it goes in exactly the wrong direction. We should be devoting more resources to understanding the *causes* and *prevention* of cancer. Even so, there’s no reason not to fight cancer on all fronts.
- Current cancer treatments make a patient a patient for the rest of their shortened lives; this method might make a patient able to live a normal, unhospitalized, non-”patient”-like life for a whole lot longer.
- Doesn’t matter. The true test is whether or not it works. Do a study, start some trials, and if we then increase survival rates and quality of life? Bring it on.

References

1. “Defining Cancer”. *National Cancer Institute*. Retrieved 10 June 2014.
2. Tarney, CM; Han, J (2014). “Postcoital bleeding: a review on etiology, diagnosis, and management.” *Obstetricsand Gynecology International* 2014: 192087. doi:10.1155/2014/192087. PMID 25045355.
3. Kumar V, Abbas AK, Fausto N, Mitchell RN (2007).*Robbins Basic Pathology* (8th ed.). Saunders Elsevier. pp. 718–721. ISBN 978-1-4160-2973-1.
4. *World Cancer Report 2014*. World Health Organization. 2014. pp. Chapter 5.12. ISBN 9283204298.
5. Dunne, EF; Park, IU (Dec 2013). “HPV and HPV associated diseases.” *Infectious Disease Clinics of North* 9 *America* 27 (4): 765–78. doi:10.1016/j.idc.2013.09.001.PMID 24275269.
6. Walker, SH; Duncan, DB (1967). “Estimation of the probability of an event as a function of several independent variables”. *Biometrika* **54**: 167–178. doi:10.2307/2333860.
7. Menard, Scott W. (2002). *Applied Logistic Regression* (2nd ed.). SAGE. ISBN 978-0-7619-2208-7.
8. Hosmer, David W.; Lemeshow, Stanley (2000). *Applied Logistic Regression* (2nd ed.). Wiley. ISBN 0-471-35632-8.
9. Hosmer, David W.; Lemeshow, Stanley (1980). *A goodness-of-fit test for the multiple Logistic Regression model”. Communication in statistics* A10:1043-1069.
10. Palei, S. K.; Das, S. K. (2009). “Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach”. *Safety Science* 47: 88–96.
11. doi:10.1016/j.ssci.2008.01.002.
12. Peduzzi, P; Concato, J; Kemper, E; Holford, TR; Feinstein, AR (December 1996). “A simulation study of the number of events per variable in logistic regression analysis.”
13. *Journal of Clinical Epidemiology* 49 (12): 1373–9. doi:10.1016/s0895-4356(96)00236-3. PMID 8970487.
14. Human Papilloma Virus ICMR: High power Committee to Evaluate Performance of ICMR, 2012–2013. New Delhi, India: ICMR; 2014. Disease Specific Documents for XII plan.