

# Anticipation of the Significance of Risk Factors in Cervical Cancer for Low Incoming Country: Bangladesh Perspective

Sayed Asaduzzaman<sup>1\*</sup>, Kawsar Ahmed<sup>1</sup>, Setu Chakraborty<sup>2</sup>, Md. Goljar Hossain<sup>3</sup>, Mamun Ibn Bashar<sup>3</sup>, Touhid Bhuiyan<sup>4</sup> and Subrata Sarker Chandan<sup>5</sup>

**Abstract**— Cervical cancer is the second alarming cancer for women of low incoming countries like Bangladesh. In future it would be the main cause of death of Bangladeshi women by cancer. To find the significant factors, association among them and making a precedence list among them by data mining and statistical approaches. During, February 2014 till July 2014 a case-control study has been acquitted on 436 participants of both patients (199) and non-patients (237). Using an accurate questionnaire based on previous study the whole data collection process done in the different part of the Dhaka cities and diagnostic center. About 10 factors like first sex at the age below 16, Lack of knowledge about cervical cancer, number of children above 3, STI (Sexually Transmitted Infection) affection, previous cervical cancer history are founded highly significant by the statistical analysis and later those factors were given precedence by data mining process Ranker algorithm with different attribute evaluator. Oral contraception taken, contraception used and vaccine taken factors are lower significant than the other factors by the analysis. Both data mining and statistical approaches depict a comparative analysis and by the result the significant factors and the significance priority can be measured.

**Index Terms**— Cervical cancer, Data mining, Statistical approach, Significant Factors, Low incoming country.

## 1 INTRODUCTION

Every year, over about 88% of women deaths occur from cervical cancer in developing country like Bangladesh, India, Pakistan etc. due to lack of knowledge, gender discrimination and extreme poverty which limit proper care to women [1]. Corruption of the highly regulated system of normal cell growth, division, and death resulted in cancer. Cervical cancer is a leading type of cancer among many other cancer like breast cancer, skin cancer, lung cancer, brain cancer, prostate cancer, colorectal cancer, stomach cancer, and other melanoma. Massive number of illiterate and conservative women of societies has no idea about cervical cancer.

Cervical cancer is associated with development of the cervix. The incidence of cervical cancer is started with a continuous process of squamous metaplasia process during puberty. This process involves transformation of columnar epithelium of the cervix into squamous epithelium, in which transitional cells support HPV replication and resulted in cervical intra-epithelial neoplasia (CIN) 2 or CIN3 lesions and, eventually, development of invasive cervical cancer. Multiple sexual partners and early sexual activity influence squamous metaplasia [2].

Five or more full term pregnancies, use of oral contraceptives for five or more years, smoking, previous exposure to sexually transmitted infections (STIs), i.e. chlamydia trachomatis, some herpes viruses and HIV are some leading factors associated with an increase in the risk of cervical cancer among HPV-DNA positive women [2]. Estrogen and its receptors promote cervical cancer because they are strongly associated with human papillomavirus (HPV) infections in combination with HPV oncogenes [3]. Cervical cancer is also regulated by lack of nutrition like vitamins [4] and aging usually after post-menopausal period at 45 or above [5]. Sometimes cervical cancer is caused by genetic factors, but this evidence is not

clear [2].

Some preventive measure of cervical cancer includes avoidance of multiple sexual partner, oral contraceptives, smoking, multiple pregnancies; avoidance of early marriage and multiple marriages; and use of condom during sexual intercourse. Cervical cancer can also be prevented by eating nutritious food which are rich in and by vaccination [6]. Surgery, hysterectomy, chemotherapy, drugs targeting estrogen and its receptors may be effective in treating and/or preventing cervical cancer.

A data mining and statistical approaches takes place in the paper. The significant patterns of the factors and finding the precedence among the significant factors have been performed using statistical software (SPSS) and Data Mining Software (WEKA).

## 2 BACKGROUND

According to National Cancer Institute, cervical cancer is a slow-growing cancer that forms in tissues of the cervix, which is an organ that connecting the uterus and vagina [7]. It is exhibits no symptoms in earlier stage but in later stage of cancer it shows symptoms like vaginal bleeding; pelvic pain and pain during sexual intercourse [8,9]. It can be detected with regular Pap test [7]. Human papillomavirus (HPV) is the actual culprit of cervical cancer which has the ability to invade other parts of the body and causing infection [10, 11].

Cervical cancer is the fourth most common cause of cancer in women. In 2012, it was estimated that there were 266,000 deaths within 528,000 cases of cervical cancer worldwide [12]. Cervical cancer is a leading cause of cancerous death in Bangladesh. 561, 583 and 574 women were affected by cervical cancer in 2005, 2006 and 2007 and it is second common malignancies (21.5%) in females of Bangladesh [13].

Hence, the actual reason and total curing procedure of can-

cer is not invented yet. Some general symptoms and risk factors of cervical cancer have been discovered by many statistical analyses. However, identification of environmental as well as genetic factors is very important in developing novel methods of cervical cancer prevention. There are lots of works to detect the risk factors of cervical cancer using population based case control study [11], several databases, and algorithm and induction techniques [14]. Some researchers tried to predict cancer risk using data mining technique [15-19]. Specifically there were no work of cervical cancer risk prediction using data mining or Statistical approaches.

### 3 METHODOLOGY

There are four parts in this section. Those are data collection, data preprocessing, statistical analysis and Data Mining approaches using WEKA. Those sections are described sequentially below.

#### 3.1 DATA COLLECTION

The data of total 436 interviewers were collected from different diagnosis center and some areas of Dhaka city. The data were collected from 436 female participants whose age range was between bellow 30 to above 60 years Some risk factors were considered for cervical cancer assessment in Bangladeshi population by former study, which includes- age, multiple sexual partner, lack of correct condom utilization, the age at first sexual intercourse of the woman, use of oral contraceptives for five or more years; high parity (five or more full term pregnancies); and previous exposure to other sexually transmitted infections (STIs), lack of proper nutrition, smoking, and sometimes it is genetic risk. A questioner was designed to collect by the former study.

#### 3.2 DATA PREPROCESSING

Data transformations, data reduction, data cleaning, data integration, data discretization are the five major tasks of the data pre-processing to avoid incomplete, inconsistent and noisy data. Conversion of the raw data to a reasonable level is the main task of data preprocessing. There is some noise in the data which were processed by some process. Incomplete data hampers the analysis which has been eliminated or leveled. A little bit of data were changes to avoid collision of the data analysis.

#### 3.3 STATISTICAL ANALYSIS

The statistical approaches were used to find the frequencies, crosstab, bar diagrams, binary logistic regression. The whole analysis was done in Statistical Package for Social Science (SPSS version 20.0). Table 1 was simulated by cross tab and frequency and Table 2 was prepared by binary logistic regression with 95% confidence interval.

#### 3.4 DATA MINING APPROACHES USING WEKA

Highly significant 10 factors have been exploited from the analysis of the statistical approaches and depending on the

results of SPSS. Then those factors are ranked by ranking algorithm with three different attribute evaluators OneRAttributeEval, ReliefAttributeEval, CorrelationAttributeEval .

### 4 RESULTS AND ANALYSIS

The statistical analysis of the results has been shown in this section. Here the frequency distribution of various factors with cervical cancer has been shown in Table 1. and Odds ratio and Confidence Interval of the associated Factors in Binary logistic regression analysis has been shown in Table 2. shows.

The frequency table was contrived by comparing the results of SPSS and WEKA. Both statistical and data mining approaches shows same frequency. Age range between 27 and 80 where the mean age was 53, approximately 436 Bangladeshi women's data were analyzed. Here 237 women were not affected (control group) and 199 women were affected by cervical cancer (case group). According to the case group, 31.7% patient's age lies between 60 and 80 and 66.8% patient's age lies between 46 and 60. Education levels of the control (affected) group participants are illiterate (55.30%), primary (27.1%), secondary (11.1%) and undergraduate/above (6.5%) severally. On the other hand in control (unaffected) group participants are illiterate (19%), primary (51.9%), and secondary (13.9%) and undergraduate/above (15.2%) respectively which has been shown in Table 1. Among 199 affected women about 128 women take adequate fruits and vegetables whether 210 of 237 unaffected women take adequate fruits and vegetables by Table 1. Among the patients 87.4% are married and the rest are married.

By the Table 1, 195 participants (98%) had no knowledge about cervical cancer among 199 affected where 75 participants (33.6%) had knowledge about cancer among 237 unaffected women. 19.6% women had more than 3 sexual partners in case (affected) group and 6.3% women had more than 3 sexual Partners in control (unaffected) group vice versa. The table also illustrates that only 5% and 6.8% women of both affected and unaffected group took cancer vaccine.

By the table it is clear that in control group most of the participants (91.9%) had their first sexual intercourse after the age of 16 years. Number of the affected women whose first sexual intercourse was before the age of 16 was about above two third of the total number.

About 95.5% affected participants had taken oral contraception when 96.5% affected participants had not used condom during sex. More than two third (73.4%) women had children among 3 to 5 in case group and one third 16.6% women had more than 5 children. On the other hand more than two third (82.7) women have 1 or 2 children in unaffected group.

The Confidence Intervals (C.I) and odds ratio with p-value and standard error of the associated factors has been shown in Table 2. Here the significantly associated factors has been shown in the table using (\*) sign. Those which factors are below the flat marginal value of  $p=0.05$ . From the Table 2 another factor has been identified Knowledge (p value is 0.000) .It means that those who has the knowledge of cancer does not have the possibility of cancer.

According to Table 2 about 10 factors have been chosen which are highly significant. Table 3 illustrates that the com-

parison of the results of Ranker algorithm by three different attribute evaluator. Ranker algorithm shows the preceding sequence of the factors. The Higher precedence depicts higher priority of the factor to be more significant. Although all the factors of Table 3 are significant but the results of the (OneR-AttributeEval, ReliefFAttributeEval, CorrelationAttributeEval) evaluators shows the higher significance and the table also gives some decisions like factor (Number of Childs) got higher precedence according the table. So it can be said that the factor (Number of Childs) is highly significant than the others.

## 5 DISCUSSION

By a registry, after breast cancer (25.6%) cervical cancer is the second (21.5%) leading cancer for Bangladeshi women. According to "National Institute of Cancer Research and Hospital" among the cervical cancer patients 8% widow, 91% were married and 97 % were housewife [13]. Among the 199 cancer patients about 87.4 % were married and 4.5% were widow which was shown through the analysis. It can be noted by the analysis that the possibility to be affected by cancer is much higher whose ages are above 46. The largest part of the cancer affected women's age was among 46 to 60. Being a developing country most of the women of Bangladesh are uneducated [13]. According to our analysis, about two third of the total cancer affected patients were uneducated. The fact is that the largest amount of cancer affected women has no idea about cancer. There are some strongly correlated factors like "whose first sex age was below 16" and "whose number of children was above 3" which were observed by the analysis. Among all the factors STI was found as highly significant factor whose p value is (0.043). According to Table 2 "Take Adequate Food", "Oral contraception", "Family cancer history" and "knowledge about cancer" were also found significant factors. By the analysis, those who took vaccine have the fewer possibility of being pretended by cervical cancer. The former study depicts that STI [11], vaccin [6], Adequete food [6], history [2], first sex age, number of sex partner, oral contraception were detected as significant which are also has been discovered by the analysis.

## 6 CONCLUSIONS

Bangladeshi women are highly affected by cervical cancer. A registry of National Cancer Institute of Bangladesh shows that cervical cancer is the second (21.5%) leading cancer after breast cancer (25.6%) in women [13]. The results of this study show that behavioral interventions which aims at promoting sexual behaviors protective of STI transmission and encourage condom use for sexual intercourse. This manner can prevent cervical cancer. Death by cervical cancer is certain due to lack of knowledge of mass people and difficulties of proper diagnosis process. In this paper, significant risk factors have been analyzed through data mining and statistical approaches. There are other approaches to detect cervical cancer risk like ANN, SVM and DT etc. SVM is not is well for discrete data and its kernel choice is more difficult for algorithm development. DT creates complex tree for categorical variables. In ANN, VC dimension is unclear what is important for good solutions. On the other hand main drawback is it can be retained. For those reasons and simplicity we chose Data mining

and Statistical Approach for analysis. An efficient approach for the extraction of significant pattern from data warehouse for efficient prediction of cervical cancer has been provided. The precedence of the significant factors of statistical analysis has been given priority sequence by data a mining approach which shows the higher risk levels of the factors.

## ACKNOWLEDGEMENT

The authors are grateful to those who gave their valuable data and who participated in this research work

## REFERENCES

- [1] Ginsburg, O. M. ; Breast and cervical cancer control in low and middle-income countries: Human rights meet sound health policy. *Jour of Cancer Pol* **2013**, *1*, e35–e41.
- [2] Shepherd, J. P.; Frampton, G. K.; Harris, P. Interventions for encouraging sexual behaviors intended to prevent cervical cancer. *Cochrane Database Syst Rev* **2011**, *4*.
- [3] Chung, S.H.; Franceschi, S.; Lambert, P. F.. Estrogen and ER $\alpha$ : Culprits in cervical cancer?. *Trends in Endocrinology & Metabolism* **2010**, *21*, 504–511.
- [4] Closas, R. G.; Castellsagué, X.; Bosch, X.; González, C. A. The role of diet and nutrition in cervical carcinogenesis: a review of recent evidence. *Inter jour of cancer* **2005**, *117*, 629–637.
- [5] Castanon, A.; Landy, R.; Cuzick, J.; Sasieni, P.. Cervical screening at age 50–64 years and the risk of cervical cancer at age 65 years and older: population-based case control study. *PLoS medicine* **2014**, *11*, e1001585.
- [6] Einstein, M. H.; Baron, M.; Levin, M. J.; Chatterjee, A.; Edwards, R. P.; Zepp, F.; Carletti, I.; Dessy, F. J.; Trofa, A. F.; Schuind, A.; others. Comparison of the immunogenicity and safety of Cervarix™ and Gardasil® human papillomavirus (HPV) cervical cancer vaccines in healthy women aged 18–45 years. *Human vaccines* **2009**, *5*, 705–719.
- [7] "Defining Cancer", National Cancer Institute, (2014).
- [8] "Cervical Cancer Treatment (PDQ®)", National Cancer Institute, (Retrieved 24 June, 2014).
- [9] Shapley, M.; Jordan, J.; Croft, P. R. . A systematic review of postcoital bleeding and risk of cervical cancer. *British jour of gen prac* **2006**, *56*, 453–460.
- [10] Schiffman, M.; Wentzensen, N.; Wacholder, S.; Kinney, W.; Gage, J. C.; Castle, P. E. . Human papillomavirus testing in the prevention of cervical cancer. *Jour of the National Cancer Ins* **2011**, 103368–383.
- [11] Johnson, G. A.; Unger, E. R.; Ouattara, E. B. Coulibaly, ; K. T.; Maurice, C.; Vernon, S. D.; Sissoko, M.; Greenberg, A. E.; Wiktor, S. Z.; Chorba, T. L. Assessing the relationship between HIV infection and cervical cancer in Cote d'Ivoire: a case-control study. *BMC infectious diseases* **2010**, *10*, 242.
- [12] World Cancer Report, 2014; World Health Organization, 2014; pp. Chapter 5.12.
- [13] Cancer Registry Report, 2005-2007; National Institute of Cancer Research and Hospital, Dhaka, Bangladesh; Published on December (2009).
- [14] Ho, S. H.; Jee, S. H.; Lee, J. E.; Park, J. S. . Analysis on risk factors for cervical cancer using induction technique, Expert Systems with Applications **2004**, *27*, 97–105.

- [15] Jesmin T, Ahmed K, Rahman MZ, Miah MBA. Brain cancer risk prediction tool using data mining. *International journal of computer applications*. 2013; 61(12): 22-27.
- [16] Ahmed K, Emran AA, Jesmin T, Mukti RF, Rahman MZ, Ahmed F. Early detection of lung cancer risk using data mining. *Asian pacific journal of cancer prevention*. 2013; 14 (1): 595–598.
- [17] Ahmed K, Jesmin T, Fatima U, Moniruzzaman M, Emran AA, Rahman MZ. Intelligent and effective diabetes risk prediction system using data mining. *Oriental Journal of Computer science & Technology*. 2012; 5 (2): 215-221.
- [18] Ahmed K, Jesmin T, Rahman MZ. Early prevention and detection of skin cancer risk using data mining. *International journal of computer applications*. 2013; 62(4):1 -6.

**Table 1.**

FREQUENCY DISTRIBUTION OF VARIOUS FACTORS

| Variable                            | Category             | Patient        |                  |
|-------------------------------------|----------------------|----------------|------------------|
|                                     |                      | Affected N (%) | Unaffected N (%) |
| Education                           | Illiterate           | 110(55.3)      | 45(19.0)         |
|                                     | Primary              | 54(27.1)       | 123(51.9)        |
|                                     | secondary            | 22(11.1)       | 33(13.9)         |
|                                     | Under graduate/above | 13(6.5)        | 36(15.2)         |
| Knowledge about cancer              | Yes                  | 4(2)           | 75(33.6)         |
|                                     | No                   | 195(98)        | 162(68.4)        |
| Ever had any cancer                 | Yes                  | 20(10.1)       | 15(6.3)          |
|                                     | No                   | 179(89.9)      | 222(93.7)        |
| Cancer vaccine taken                | Yes                  | 10(5)          | 16(6.8)          |
|                                     | No                   | 189(95.0)      | 221(93.2)        |
| Take Adequate Fruits and vegetables | Yes                  | 128(64.3)      | 210(88.6)        |
|                                     | No                   | 71(35.7)       | 27(11.4)         |
| Number of Sex partner               | 1-2                  | 160(80.4)      | 222(93.7)        |
|                                     | 3+                   | 39(19.6)       | 15(6.3)          |
| First sex age                       | Below 16             | 146(73.4)      | 21(8.9)          |
|                                     | Above 16             | 53(26.6)       | 216(91.9)        |
| Contraception used                  | Yes                  | 7(3.5)         | 163(68.8)        |
|                                     | No                   | 192(96.5)      | 74(31.2)         |
| Oral contraception taken            | Yes                  | 190(95.5)      | 74(31.2)         |
|                                     | No                   | 9(4.5)         | 163(68.8)        |
| Number of Childs                    | 1-2                  | 20(10.1)       | 196(82.7)        |
|                                     | 3-5                  | 146(73.4)      | 26(11.0)         |
|                                     | 5+                   | 33(16.6)       | 15(6.3)          |
| Affected by STI                     | Yes                  | 36(18.1)       | 4(1.7)           |
|                                     | No                   | 163(81.9)      | 233(98.3)        |
| Age                                 | Below 30             | 0(0.0)         | 3(1.3)           |
|                                     | 31-45                | 3(1.5)         | 5(2.1)           |
|                                     | 46-60                | 133(66.8)      | 134(56.5)        |
|                                     | Above 60             | 63(31.7)       | 95(40.1)         |
| Family cancer history               | Yes                  | 55(27.6)       | 18(7.6)          |
|                                     | No                   | 144(72.4)      | 219(92.4)        |
| Marital status                      | Divorce              | 6(3.0)         | 0(0.0)           |
|                                     | Married              | 174(87.4)      | 236(99.6)        |

|  |                   |                   |                  |
|--|-------------------|-------------------|------------------|
|  | Separate<br>Widow | 10(5.0)<br>9(4.5) | 1(0.4)<br>0(0.0) |
|--|-------------------|-------------------|------------------|

**Table 2.**

ODDS RATIO AND CONFIDENCE INTERVAL WITH P VALUE OF THE FACTORS

| Factors   | B              | S.E.          | Sig. (p)                | Exp.<br>(B)      | 95% C.I. for EXP(B) |                  |
|---|----------------|---------------|-------------------------|------------------|---------------------|------------------|
|   |                |               |                         |                  | Lower               | Upper            |
| Premature chronicle of cancer<br>No(ref)<br>Yes                     | -1.148         | 1.316         | 0.383                   | 0.317            | 0.024               | 4.19             |
| Cancer vaccine taken<br>No(ref)<br>Yes                              | -1.519         | 1.545         | 0.326                   | 0.219            | 0.011               | 4.527            |
| Take Adequate Food(*)<br>No(ref)<br>Yes                             | -2.475         | .835          | 0.003                   | 0.084            | 0.016               | 0.433            |
| How many Sex partner<br>1-2(ref)<br>3+                              | 1.129          | 1.155         | 0.328                   | 3.094            | 0.322               | 29.755           |
| First sex (*)<br>Below 16(ref)<br>Above 16                          | -2.083         | 0.533         | 0.000                   | 0.124            | 0.044               | 0.354            |
| Oral contraception taken<br>No(ref)<br>Yes                          | 1.709          | 3.295         | 0.604                   | 5.523            | .009                | 3520.785         |
| Contraception used<br>No(ref)<br>Yes                                | -3.308         | 3.278         | 0.313                   | 0.037            | 0.00                | 22.555           |
| Number of Childs (*)<br>1-2(ref)<br>3-5<br>Above 5                  | 3.114<br>2.414 | .617<br>1.005 | 0.000<br>0.000<br>0.016 | 22.517<br>11.180 | 6.724<br>1.560      | 75.409<br>80.131 |
| Previous Cancer history of<br>cervical cancer (*)<br>No(ref)<br>Yes | 2.743          | 2.743         | .006                    | 15.539           | 2.188               | 110.350          |
| Affected by STI (*)<br>No(ref)<br>Yes                               | 2.564          | 2.564         | 0.043                   | 12.984           | 1.079               | 156.201          |
| Knowledge(*)<br>No(ref)<br>Yes                                      | -5.779         | 1.115         | 0.00                    | 0.003            | 0.000               | 0.028            |

# IJSER