STATISTICAL MODELING OF TOBACCO RELATED CANCER IN KAMRUP URBAN DISTRICT

Surobhi Deka[1,] * and Sarita Barman[2]
[1] Department of Statistics, Cotton University, Guwahati- 781001, Assam
*Corresponding Author: surobhi.deka@cottonuniversity.ac.in

Abstract

The purpose of the study was to highlight the incidence and pattern of Tobacco Related Cancer (TRC) in resident population of Kamrup Urban District (KUD) for the year 2012-2014. Four continuous probability distributions viz. Normal distribution, Log-Normal distribution, Erlang distribution and Weibull distribution were considered to find the best fitting probability distribution for tobacco related cancer incidence. The cumulative distribution functions for the general form of the continuous probability distributions were tested for goodness of fit by using the Kolmogorov-Smirnov test. Graphical plots of theoretical and observed cumulative distribution functions were applied to confirm the best fit for the mentioned probability distributions. Finally, the goodness of fit test results were compared and the Weibull distribution has been proved to be the most appropriate distribution for describing the tobacco related cancer incidence data in Kamrup Urban District for the year 2012- 2-14.

Key words: Tobacco Related Cancer, continuous probability distributions, best fit.

1. INTRODUCTION

Cancer is one of the leading causes of morbidity and mortality worldwide despite sophistication in diagnosis and advances in treatment. Tobacco use is a global epidemic among young people and it is one of the leading causes of cancer worldwide. One-half of adult smokers die prematurely from tobacco-related diseases.

About 40% of all cancers are tobaccos related and 90% of the oral cancers are due to use of tobacco. Tobacco use accounts for nearly half of all cancers among males and a quarter of all cancer among females. The anatomical sites of cancer associated with the use of Tobacco are Lip, Tongue, Mouth, Oropharynx, Hypopharynx, Pharynx, Esophagus, Larynx, Lung and Urinary Bladder. The incidence and relative proportion of specific sites of cancer associated with the use of tobacco varies according to the type of tobacco and the manner of its consumption (Kathirvel et al. (2014)). Tobacco is consumed in myriad forms in India which includes smoking as well as smokeless tobacco. Tobacco is smoked, chewed, sniffed, and kept in the oral cavity and reverse smoking. Smoked tobacco mainly consists of cigarettes and beedis. Smokeless tobacco includes tobacco that is chewed with or without betel nut, khaini and moist oral snuff. It contains 4000 different chemicals and more than 60 of these are carcinogenic (Rani et al. (2003), Mathur and Shah (2011)). Bidi is the most popular prevalent smoking product consumed in rural areas (John (2005), Gupta et al. (2010)) in comparison to cigarette smoking more preferably used in urban areas (Gupta et al. (2010)). Hookah, chuttas, dhumti, chillum, cigars, cheroots and pipes are some other forms of tobacco smoking in different parts of India (Jindal et al. (2006)). Dry tobacco, such as paan masala, gutkha and mawa are also popular in many parts of India (Rooban et al. (2010)). Furthermore, oral tobacco such as mishri, gul, gudakhu are widely used as topical applications on teeth and gums (Rani et al. (2003)). Smokeless tobacco is consumed predominantly by chewing in form of pan (piper betel

leaf filled with sliced areca nut, lime, catechu, and other spices chewed with or without tobacco), pan-masala or gutkha (a chewable tobacco containing areca nut) and mishri (a powdered tobacco rubbed on the gums as toothpaste) (Gupta and Ray (2003), Dobe et al. (2006), Gupta (2013)).The number of new cases of all cancer was increased from 155.3 to 188.5 and 102.7 to 165.3 per 100,000 men and women respectively from the year 2007 to 2011 in Kamrup Urban District (KUD) (Sharma, et al. (2016)*).* The objective of this study is to highlight the incidence and pattern of Tobacco Related Cancers with respect to all other sites of cancer in Kamrup Urban District (KUD). The pattern and incidence of tobacco related cancers in resident population of KUD are reported for the year 2012-2014. In order to describe the behavior of cancer incidence at the study area, it is necessary to identify the distribution(s) which best fit the data. In this study, four continuous probability distributions viz. Normal distribution, Log-Normal distribution, Erlang distribution and Weibull distribution are considered to find the best fitting probability distribution of tobacco related cancer incidence.

2. DATA AND METHODOLOGY

2.1 Data

Kamrup Metropolitan is one of the 35 districts in Assam state in north-eastern India. The district occupies an area of 1527.84 km². Kamrup metropolitan district is located between 25°43′-26°51′N Latitude and 90°36′–92°12′E Longitude. The district is bounded on the West and North by the Kamrup district and on the East by the Morigaon district. On the South, lies the state of Meghalaya. According to the 2011 census, Kamrup Metropolitan district has a population of 1,260,419 (Census 2011). The combined 3- year 2012–2014 population distribution by 5-year age group and gender is shown in Figure 2. 1. The district has a population density of 2,010

inhabitants per square kilometer (5,200/sq mi). Its population growth rate over the decade 2001-2011 was 18.95%. Kamrup Metropolitan has a sex ratio of 922 females for every 1000 males and a literacy rate of 88.66%. The data regarding Tobacco Related Cancer Incidence has been collected from the Dr B. Borooah Cancer Institute, Guwahati from its latest available three-year report of Kamrup urban district cancer incidence data for the period 2012-2014. Population Based Cancer Registry (PBCR) was set up in the Department of Pathology at Dr. B. Borooah Cancer Institute (BBCI), Guwahati, in 2003 to generate authentic and reliable data on cancer incidence and mortality pattern of Kamrup urban district (KUD) of Kamrup district. PBCR-Guwahati covers an area of 267.1 km^2.

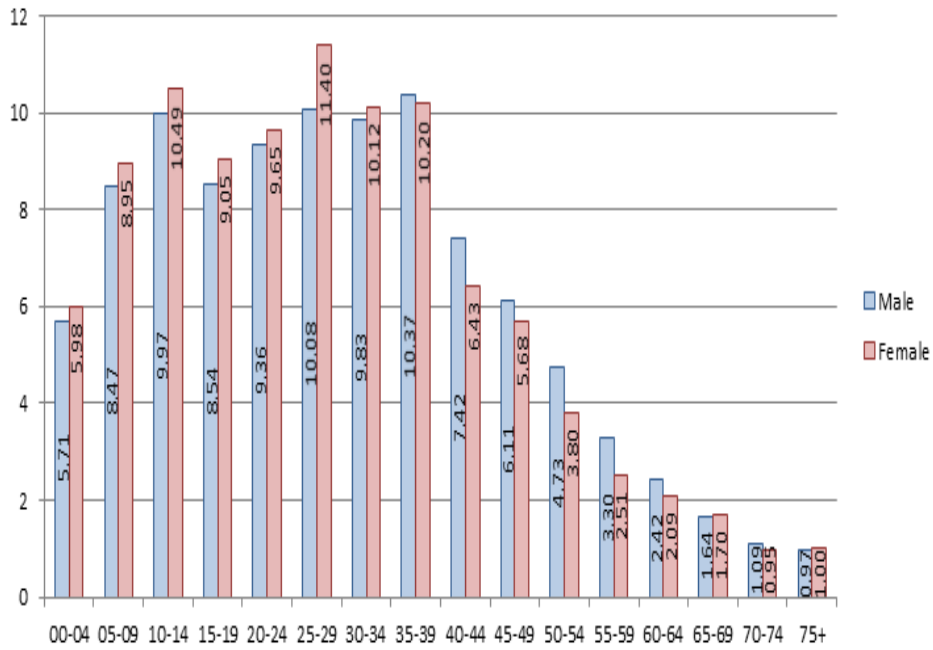


Figure 2.1: Population distribution by 5-year age group and gender.

Table 2.1a: Number of Cancer Incidence of Males by 5-year age group

ICD-10 Code	Site	00-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75+	Total	%
C00	Lip	-	-	-	-	1	2	1	2	-	-	1	1	-	8	0.52
C01-C02	Tongue	-	1	2	2	3	11	10	13	16	22	4	11	14	109	7.15
C03-C06	Mouth	-	1	1	4	5	11	12	20	22	49	12	14	10	161	10.56
C10	Oth. Oropharynx	-	-	-	-	3	2	10	9	17	11	4	5	7	68	4.46
C12-C13	Hypopharynx	-	-	2	6	10	20	31	38	42	40	30	24	30	273	17.90
C14	Pharynx Unspecified	-	-	-	-	1	-	3	6	10	5	6	7	12	50	3.28
C15	Oesophagus	-	-	1	7	15	23	45	65	66	62	65	43	47	439	28.79
C32	Larynx	-	-	-	1	3	8	15	11	14	16	17	15	9	109	7.15
C33-C34	Lung etc.	-	1	1	1	4	6	16	26	40	45	25	47	45	257	16.85
C67	Urinary Bladder	-	-	-	-	-	3	3	3	5	16	6	9	6	51	3.34
	All TRC Sites	-	3	7	21	45	86	146	193	232	266	170	176	180	1525	100.0

In KUD, the relative proportion of cancers associated with the use of tobacco for male and female is 49.7% and 24.1% respectively when compared to all sites. The number and relative proportion of cancers associated with the use of tobacco has been worked out according to the monograph of the International Agency for Research on Cancer (IARC 1987). The anatomical sites of cancer that have been associated with the use of tobacco (TRC) include lip, tongue, mouth, pharynx (oropharynx and hypopharynx), oesophagus, larynx, lung and urinary bladder. Table 2.1a and Table 2.1b provide the

number and relative proportion of sites of cancer associated with the use of tobacco in KUD. Esophageal cancer alone

contributes 28.79% in males to all TRC cases, followed by hypopharynx cancer 17.90% of cases. This is diagrammatically given in Figure 2.1a.

Table 2.1b: Number of Cancer Incidence of Females by 5-year age group.

CD-10 Code	Site	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75+	Total	%
C00	Lip	-	-	-	-	-	-	1	-	-	1	-	1	-	2	-	5	0.87
C01-C02	Tongue	-	-	-	-	1	3	3	2	3	2	2	7	9	4	5	41	7.12
C03-C06	Mouth	-	-	-	2	3	2	3	3	8	9	8	12	5	13	15	83	14.41
C10	Oth. Oropharynx	-	-	-	-	2	-	1	-	2	3	3	2	3	3	4	23	3.99
C12-C13	Hypopharynx	1	-	-	-	-	-	2	3	8	5	3	5	10	4	1	42	7.29
C14	Pharynx Unspecified	-	-	-	-	-	-	1	-	3	1	-	-	-	1	4	10	1.74
C15	Oesophagus	-	-	-	1	1	4	1	13	21	30	40	42	33	32	27	245	42.53
C32	Larynx	-	-	-	-	-	-	-	1	2	1	6	3	2	-	1	16	2.78
C33-C34	Lung etc.	-	-	-	-	1	4	5	6	13	10	22	18	9	9	8	105	18.23
C67	Urinary Bladder	-	-	-	-	-	-	-	1	-	-	-	1	2	2	-	6	1.04
	All TRC Sites	1	-	-	3	8	13	17	29	60	62	84	91	73	70	65	576	100.00

In female, the highest contributor is also esophageal cancer contributing a total of 42.53% to all TRC cases, followed by lung cancer 18.23% and mouth cancer with

14.41% of cases. Figure 2.1b diagrammatically illustrates the proportion of specific tobacco related sites for female. Also, we can illustrate that carcinoma of lip, mouth, esophagus, and lung was high in females compared to males, while it is also observed that carcinoma of tongue, oropharynx, hypopharynx, pharynx, larynx, and bladder is high in males compared to females. This comparison showed that all the men are more at risk of pharyngeal carcinoma compared to female; this difference may be associated with the form of tobacco use.

Figure 2.1a: Incident Cancer of Males

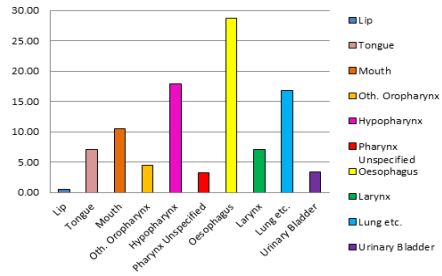
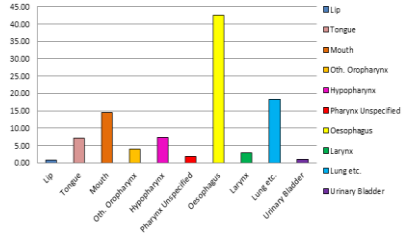


Figure 2.1b: Incident Cancer of Females



2.2 Probability Distributions

In this study, the two parameter continuous probability distributions namely – Normal Distribution, Log-Normal Distribution, Erlang Distribution and Weibull Distribution is considered to find the best fitting probability distribution function of Tobacco Related Cancer Incidence.

Normal Distribution: A random variable X is said to have a normal distribution with parameters μ (called mean) and σ² (called variance) if its p.d.f is given by –

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}},\ -\infty<x,\mu<\infty,\ \sigma>0,$$

where the parameters of the distribution i.e., μ and σ² can be evaluated using the relationship

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\mu})^2.$$

Log Normal Distribution: The p. d. f. of the lognormal distribution is given by

$$y = f(x/\mu,\sigma) = \frac{1}{\sigma x\sqrt{2\pi}} e^{\frac{-(z-\mu)^2}{2\sigma^2}}, x > 0,$$

where $z = log\ x$ and the parameters of the distribution i.e., μ and σ can be evaluated using the following relationships

$$\hat{\mu} = \bar{z},\ \hat{\sigma} = [n^{-1}\sum_{j=1}^{n}(z_j - \bar{z})]^{\frac{1}{2}}.$$

Erlang Distribution: The Erlang distribution is a two parameter family of continuous probability distributions with p.d.f given by

$$f(x) = \begin{cases} \dfrac{\lambda^k x^{k-1} e^{-k\lambda x}}{\Gamma(k)}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

where k and λ are shape and scale parameters of the distribution and can be evaluated by solving the equations

$$\frac{1}{n}\sum_{j=1}^{n}\log x_j = \log\frac{1}{\hat{\lambda}} + \Psi(\hat{k}),\ \bar{x} = \frac{\hat{k}}{\hat{\lambda}} \text{ and } \Psi(k) = \frac{\partial \log(\Gamma(\hat{k}))}{\delta\hat{k}}.$$

Weibull Distribution: The probability density function of the two parameter Weibull Distribution is given by

$$y = f(x;\alpha,\beta) = \alpha\beta \; x^{\beta-1} e^{-\alpha x^{\beta}}, \qquad x > 0.$$

where α be the scale parameter and β be the shape parameter of the distribution. The maximum likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$ of α and β respectively satisfy the following equations

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i^{\hat{\beta}} \log x_i}{\sum_{i=1}^{n} x_i^{\hat{\beta}}} - \frac{\sum_{i=1}^{n} \log x_i}{n}, \qquad \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} x_i^{\hat{\beta}}}.$$

2.3 Test for goodness of fit -

The test applied for judging the goodness of fit of the distributions for Tobacco Related Cancer Incidence is the KOLMOGOROV-SMIRNOV TEST. The cumulative distribution functions for the general forms of the following continuous probability distributions namely Normal, Log Normal, Erlang and Weibull; were tested for goodness of fit. Massey showed that Kolmogorov-Smirnov test treats individual observation separately leading to no loss of information in grouping while loss of information in Chi-square procedure is large. Pal (1998) mentioned that the Chi-square test's sensitivity to very small cell frequencies make itself unsuitable when expected frequencies work out at less than 5 in 20 per cent of the cells. In the present case it is found that more than 50% of the cell frequencies are less than 5. Also according to Keeping (1962), Kolmogorov Smirnov test can be applied in situations where the theoretical distribution function is continuous.

The (K-S) test is a simple non-parametric test for testing whether there is a significant difference between an observed frequency distribution and theoretical frequency distribution or not. Let a random sample $X_1, X_2, ..., X_n$ be drawn from a population with unknown cumulative distribution function *F(x)* and the null

hypothesis to be tested in this case is $H_0 : F(x) = F_0(x)$ against the alternative hypothesis $H_1 : F(x) \neq F_0(x)$, $F_0(x)$ is the specified distribution function. Now, let us define the empirical distribution function as $S_n(x)$ of the ordered sample values $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ as

$$S_n(x) = \begin{cases} 0 & if \quad x \leq x_1, \\ \frac{k}{n} & if \quad x_k \leq x \leq x_{k+1}, \\ 1 & if \quad x \geq x_n. \end{cases}$$

The (K-S) test is based on the Glivenko-Cantelli theorem which states that the step function $S_n(x)$ with jumps occurring at the values of the ordered statistics for the sample approaches the true distribution for all *x*. Therefore for large *n*, the deviation between the true c.d.f. $F_0(x)$ under H_0 and the empirical distribution function $S_n(x)$ should be small for all values of *x* since the actual numerical difference $| S_n(x) - F_0(x) |$ depends on *x*, the (K-S) statistic is taken to be the supremum of such difference, say, $D_n = \sup | S_n(x) - F_0(x) |$, where D_n is known as the (K-S) statistic. Under H, the statistic D_n has a distribution which is independent of the c.d.f. *F(x)* that defines H_0. This statistic D_n is distribution free. To decide about H_0, the test criterion is to reject H_0 if D_n exceeds the tabulated value for given *n* with a pre-fixed significant level α.

We applied graphical plots of theoretical and observed cumulative distribution functions to confirm the best fit for the mentioned probability distribution.

3. DISTRIBUTION FITTING

In the present study, four continuous distributions namely Normal, Lognormal, Erlang and Weibull are considered as the probability distribution functions for Tobacco related cancer incidents. The parameters for each distribution are estimated using the maximum likelihood method for each sites of tobacco related cancer. The results are provided in Tables 3.1(a) to 3.1(j) (for males) and Tables 3.2(a) to 3.2(j) (for females). The tables also include the observed frequencies and expected frequencies obtained from the different fitted distributions. The values of Kolmogorov- Smirnov D statistics is also provided as evidence in support of goodness of fit. In order to confirm the goodness of fit for the above four distributions, we additionally applied graphical plots of theoretical and observed cumulative distribution functions (Fig 3.1, 3.2). All the computations have been carried out in the workstation MATLAB 7.0 and MS Excel.

Table 3.1 (a): Fitting of probability distributions for "OESOPHAGUS CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=56.06 σ=10.54	Lognormal μ=4.00 σ=0.20	Erlang k=24.5544 λ=0.0180	Weibull α=59.9185 β=6.1012
Below 25	0	0	0	0	1
25-30	2	1	0	0	2
30-35	6	4	3	5	5
35-40	10	10	11	12	11
40-45	20	20	24	24	24
45-50	31	33	39	36	30
50-55	38	49	49	49	39
55-60	42	49	49	49	49
60-65	40	38	31	32	42
65-70	30	26	24	24	29
70-75	24	14	13	12	12
Kolmogorov Smirnov D Statistics		0.07	0.11	0.0665	0.0266

Table 3.1 (b): Fitting of probability distributions for "HYPOPHARYNX CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=57.45 σ=10.14	Lognormal μ=4.03 σ=0.19	Erlang k=27.9094 λ=0.0175	Weibull α=61.2893 β=6.6164
Below 25	0	0	0	0	0
25-30	1	1	0	0	4
30-35	7	4	2	4	8
35-40	15	11	12	12	12
40-45	23	26	39	31	24
45-50	45	48	63	55	43
50-55	65	78	78	78	63
55-60	66	78	78	78	78
60-65	62	68	57	55	78
65-70	65	51	39	39	55
70-75	43	26	24	39	27
Kolmogorov Smirnov D Statistics		0.08	0.13	0.0691	0.0513

Table 3.1 (c): Fitting of probability distributions for "LUNG CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=60.47 σ=9.78	Lognormal μ=4.08 σ=0.18	Erlang k=31.5753 λ=0.0167	Weibull α=64.0121 β=7.4521
Below 20	0	0	0	0	0
20-25	1	0	0	0	0
25-30	1	0	0	0	2
30-35	1	1	0	0	2
35-40	4	3	3	4	4
40-45	6	8	10	11	8
45-50	16	21	23	21	17
50-55	26	31	42	42	28
55-60	40	42	42	42	42
60-65	45	42	42	42	45
65-70	25	42	27	28	42
70-75	47	20	21	21	21
Kolmogorov Smirnov D Statistics		0.13	0.129	0.0871	0.0873

Table 3.1 (d): Fitting of probability distributions for "MOUTH CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=57.00 σ=10.40	Lognormal μ=4.02 σ=0.21	Erlang k=25.0718 λ=0.0177	Weibull α=60.5669 β=6.7523
Below 20	0	0	0	0	0
20-25	1	0	0	0	0
25-30	1	1	0	0	2
30-35	4	2	2	2	3
35-40	5	5	6	6	5
40-45	11	15	15	15	11
45-50	12	20	23	23	17
50-55	20	26	30	30	26
55-60	22	30	25	30	30
60-65	49	25	20	21	30
65-70	12	17	15	15	20
70-75	14	10	15	9	9
Kolmogorov Smirnov D Statistics		0.15	0.16	0.1363	0.0988

Table 3.1 (e): Fitting of probability distributions for "TONGUE CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=55.18 σ=11.33	Lognormal μ=3.98 σ=0.23	Erlang k=19.7563 λ=0.0183	Weibull α=59.3048 β=5.6744
Below 20	0	0	0	0	0
20-25	1	0	0	0	0
25-30	2	1	1	1	1
30-35	2	2	2	3	3
35-40	3	5	10	7	5
40-45	11	9	12	11	10
45-50	10	12	14	14	12
50-55	13	19	19	19	15
55-60	16	19	14	14	19
60-65	22	14	11	11	15
65-70	4	9	10	10	10
70-75	11	5	5	5	5
Kolmogorov Smirnov D Statistics		0.10	0.16	0.097	0.0464

Table 3.1 (f): Fitting of probability distributions for "LARYNX CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=58.15 σ=10.41	Lognormal μ=4.04 σ=0.19	Erlang k=28.0955 λ=0.0173	Weibull α=61.9886 β=6.5494
Below 30	0	0	0	0	1
30-35	1	1	0	1	1
35-40	3	3	3	3	3
40-45	8	6	8	10	6
45-50	15	11	14	13	10
50-55	11	16	20	20	15
55-60	14	20	20	20	20
60-65	16	20	15	15	20
65-70	17	12	10	11	14
70-75	15	10	10	7	10
Kolmogorov Smirnov D Statistics		0.10	0.13	0.0883	0.0645

Table 3.1 (g): Fitting of probability distributions for "OROPHARYNX CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=56.43 σ=8.69	Lognormal μ=4.02 σ=0.16	Erlang k=37.1830 λ=0.0179	Weibull α=59.7256 β=6.9262
Below 35	0	0	0	0	1
35-40	3	1	1	2	2
40-45	2	4	6	6	4
45-50	10	8	10	10	7
50-55	9	13	13	13	11
55-60	17	14	13	13	13
60-65	11	12	9	9	12
65-70	4	6	5	6	7
70-75	5	3	3	2	2
Kolmogorov Smirnov D Statistics		0.031	0.10	0.0918	0.0513

Table 3.1 (h): Fitting of probability distributions for "LIP CANCER" for males (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=51.875 σ=11.575	Lognormal μ=3.92 σ=0.216	Erlang k=22.8448 λ=0.0193	Weibull α=56.4395 β=5.0107
Below 35	0	1	0	0	0
35-40	1	1	1	1	1
40-45	2	1	1	1	1
45-50	1	1	2	2	1
50-55	2	1	1	2	2
55-60	0	1	1	1	1
60-65	0	1	1	1	1
65-70	1	1	0	0	1
70-75	1	0	0	0	0
Kolmogorov Smirnov D Statistics		0.17	0.15	0.1625	0.2154

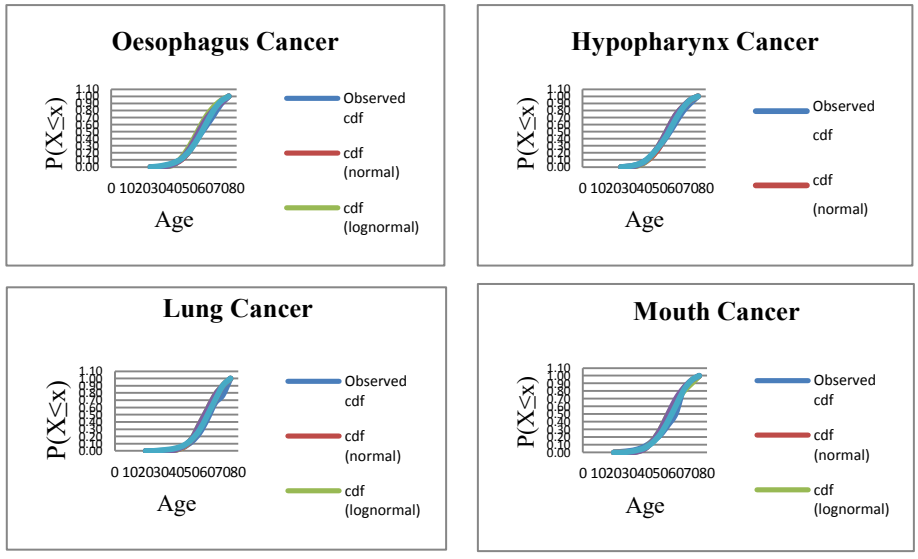
Table 3.1 (i): Fitting of probability distributions for "URINARY BLADDER CANCER" for males (2012-14)

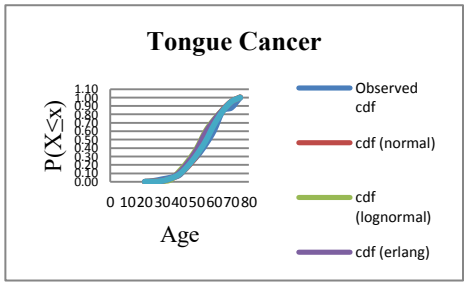
Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=61.61 σ=8.58	Lognormal μ=4.11 σ=0.15	Erlang k=44.0788 λ=0.0164	Weibull α=64.7974 β=8.6545
Below 40	0	0	0	0	1
40-45	3	1	1	1	1
45-50	3	5	5	5	3
50-55	3	6	9	7	5
55-60	5	9	10	9	9
60-65	16	10	9	9	11
65-70	6	9	7	9	10
70-75	9	5	5	5	5
Kolmogorov Smirnov D Statistics		0.14	0.22	0.1607	0.0918

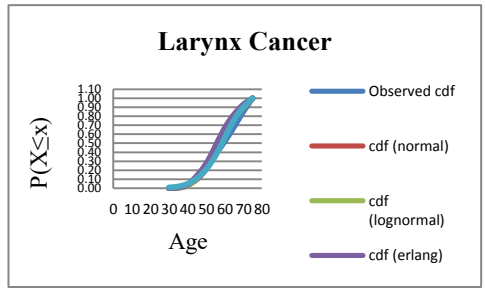
Table 3.1 (j): Fitting of probability distributions for "PHARYNX CANCER" for males (2012-14)

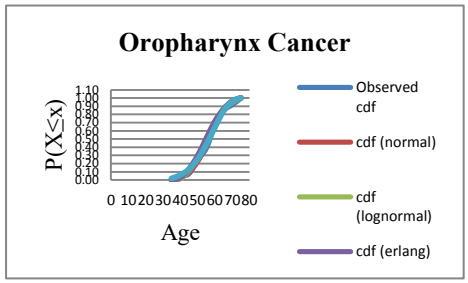
Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=60.39 σ=8.63	Lognormal μ=4.09 σ=0.15	Erlang k=44.3119 λ=0.0167	Weibull α=63.6237 β=7.8850
Below 35	0	0	0	0	0
35-40	1	0	0	0	1
40-45	0	1	1	1	2
45-50	3	4	4	4	3
50-55	6	6	8	8	5
55-60	10	8	8	8	8
60-65	5	8	8	8	8
65-70	6	6	5	5	8
70-75	7	4	4	4	4
Kolmogorov Smirnov D Statistics		0.07	0.09	0.0673	0.0627

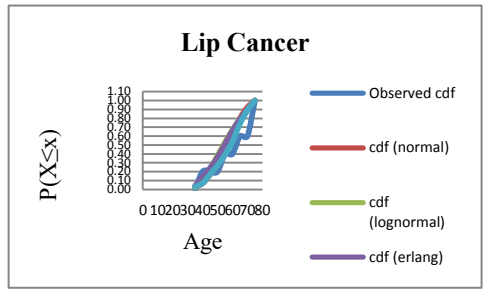
Fig.3.1 Curve for fitted and observed probability distribution functions for males (2012-14)

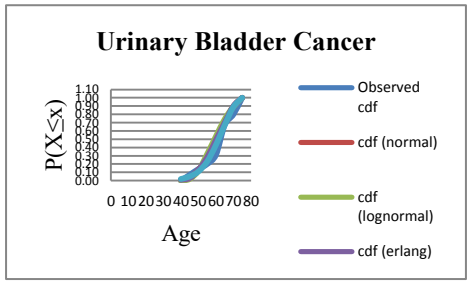



Tongue Cancer
P(X≤x)
1.10 1.00 0.90 0.80 0.70 0.60 0.50 0.40 0.30 0.20 0.10 0.00
0 10 20 30 40 50 60 70 80
Age
Observed cdf
cdf (normal)
cdf (lognormal)
cdf (erlang)


Larynx Cancer
P(X≤x)
1.10 1.00 0.90 0.80 0.70 0.60 0.50 0.40 0.30 0.20 0.10 0.00
0 10 20 30 40 50 60 70 80
Age
Observed cdf
cdf (normal)
cdf (lognormal)
cdf (erlang)


Oropharynx Cancer
P(X≤x)
1.10 1.00 0.90 0.80 0.70 0.60 0.50 0.40 0.30 0.20 0.10 0.00
0 10 20 30 40 50 60 70 80
Age
Observed cdf
cdf (normal)
cdf (lognormal)
cdf (erlang)


Lip Cancer
P(X≤x)
1.10 1.00 0.90 0.80 0.70 0.60 0.50 0.40 0.30 0.20 0.10 0.00
0 10 20 30 40 50 60 70 80
Age
Observed cdf
cdf (normal)
cdf (lognormal)
cdf (erlang)


Urinary Bladder Cancer
P(X≤x)
1.10 1.00 0.90 0.80 0.70 0.60 0.50 0.40 0.30 0.20 0.10 0.00
0 10 20 30 40 50 60 70 80
Age
Observed cdf
cdf (normal)
cdf (lognormal)
cdf (erlang)

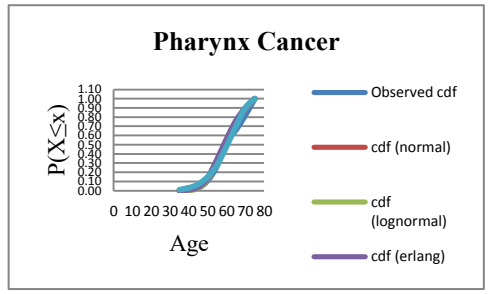

Pharynx Cancer
P(X≤x)
1.10 1.00 0.90 0.80 0.70 0.60 0.50 0.40 0.30 0.20 0.10 0.00
0 10 20 30 40 50 60 70 80
Age
Observed cdf
cdf (normal)
cdf (lognormal)
cdf (erlang)

Table 3.2 (a): Fitting of probability distributions for "OESOPHAGUS CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=58.78 σ=10.04	Lognormal μ=4.05 σ=0.19	Erlang k=29.0327 λ= 0.0171	Weibull α=62.4285 β=7.0355
Below 20	0	0	0	0	0
20-25	1	0	0	0	0
25-30	1	0	0	0	0
30-35	4	1	1	2	2
35-40	1	5	7	4	7
40-45	13	12	22	15	11
45-50	21	23	29	28	22
50-55	30	44	44	44	33
55-60	40	43	44	44	44
60-65	42	44	33	35	44
65-70	33	30	24	24	35
70-75	32	17	15	22	22
Kolmogorov Smirnov D Statistics		0.09	0.16	0.0849	0.0433

Table 3.2 (b): Fitting of probability distributions for "LUNG CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=55.59 σ=10.80	Lognormal μ=3.99 σ=0.21	Erlang k=22.7113 λ=0.0181	Weibull α=59.5797 β=6.0650
Below 25	0	0	0	0	0
25-30	1	1	0	5	1
30-35	4	2	2	2	2
35-40	5	4	5	6	5
40-45	6	10	11	11	10
45-50	13	13	19	15	12
50-55	10	17	19	16	16
55-60	22	19	15	16	19
60-65	18	15	12	13	16
65-70	9	10	8	10	11
70-75	9	6	5	5	5
Kolmogorov Smirnov D Statistics		0.09	0.19	0.1205	0.0597

Table 3.2 (c): Fitting of probability distributions for "MOUTH CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=55.15 σ=13.70	Lognormal μ=3.98 σ=0.29	Erlang k=12.8689 λ= 0.0182	Weibull α=60.1640 β=4.8710
Below 20	0	0	0	0	1
20-25	2	1	0	0	1
25-30	3	1	2	1	1
30-35	2	2	4	7	3
35-40	3	7	7	7	4
40-45	3	7	8	7	7
45-50	8	8	9	9	8
50-55	9	10	9	9	10
55-60	8	10	8	8	10
60-65	12	9	7	7	10
65-70	5	7	7	7	7
70-75	13	7	7	7	7
Kolmogorov Smirnov D Statistics		0.11	0.1340	0.1031	0.0672

Table 3.2 (d): Fitting of probability distributions for "HYPOPHARYNX CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=56.16 σ=12.93	Lognormal μ=3.98 σ=0.36	Erlang k=9.2270 λ= 0.0181	Weibull α=59.9169 β=4.9971
Below 5	0	0	0	0	0
05-10	1	0	0	0	0
10-15	0	0	0	0	0
15-20	0	0	1	0	0
20-25	0	0	1	1	0
25-30	0	1	2	2	1
30-35	0	1	4	4	2
35-40	2	3	4	4	2
40-45	3	4	4	5	4
45-50	8	5	5	5	5
50-55	5	6	4	5	6
55-60	3	6	4	5	6
60-65	5	6	4	4	6
65-70	10	4	4	4	5
70-75	4	4	4	4	4
Kolmogorov Smirnov D Statistics		0.13	0.25	0.1692	0.1274

Table 3.2 (e): Fitting of probability distributions for "TONGUE CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=56.11 σ=13.62	Lognormal μ=3.99 σ=0.28	Erlang k=13.6548 λ= 0.0180	Weibull α=60.6957 β=5.0457
Below 25	0	0	0	0	0
25-30	1	1	1	1	1
30-35	3	1	2	2	1
35-40	3	4	4	4	2
40-45	2	4	4	4	4
45-50	3	4	5	5	4
50-55	2	5	5	5	5
55-60	2	5	5	5	6
60-65	7	5	4	4	5
65-70	9	4	4	4	4
70-75	4	4	4	4	4
Kolmogorov Smirnov D Statistics		0.21	0.25	0.1993	0.1607

Table 3.2 (f): Fitting of probability distributions for "LIP CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=59.5 σ=13.27	Lognormal μ=4.054 σ=0.25	Erlang k=17.2007 λ=0.0169	Weibull α=64.4035 β=5.5940
Below 35	0	0	0	0	0
35-40	1	0	0	1	0
40-45	0	1	1	1	1
45-50	0	0	1	1	1
50-55	1	1	1	1	1
55-60	0	1	1	1	1
60-65	1	1	1	1	1
65-70	0	1	0	1	1
70-75	2	1	1	1	1
Kolmogorov Smirnov D Statistics		0.29	0.20	0.1865	0.1969

Table 3.2 (g): Fitting of probability distributions for "OROPHARYNX CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=55 .92 σ=13.48	Lognormal μ=3.99 σ=0.29	Erlang k=13.0921 λ=0.0180	Weibull α=60.7345 β=4.9844
Below 25	0	0	0	0	0
25-30	2	0	0	0	0
30-35	0	1	2	1	1
35-40	1	2	2	2	1
40-45	0	2	2	2	2
45-50	2	2	3	3	2
50-55	3	3	3	3	3
55-60	3	3	2	2	3
60-65	2	2	2	2	3
65-70	3	2	2	2	2
70-75	3	2	2	2	2
Kolmogorov Smirnov D Statistics		0.103	0.20	0.1186	0.0807

Table 3.2 (h): Fitting of probability distributions for "URINARY BLADDER CANCER" for females (2012-14)

Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=64.17 σ=10.27	Lognormal μ=4.14 σ=0.19	Erlang k=29.1408 λ= 0.0157	Weibull α=67.6568 β=9.0909
Below 40	0	0	0	0	0
40-45	1	0	0	0	0
45-50	0	1	1	1	0
50-55	0	1	1	1	1
55-60	0	1	1	1	1
60-65	1	1	1	1	1
65-70	2	1	1	1	2
70-75	2	1	1	1	1
Kolmogorov Smirnov D Statistics		0.30	0.33	0.2294	0.1608

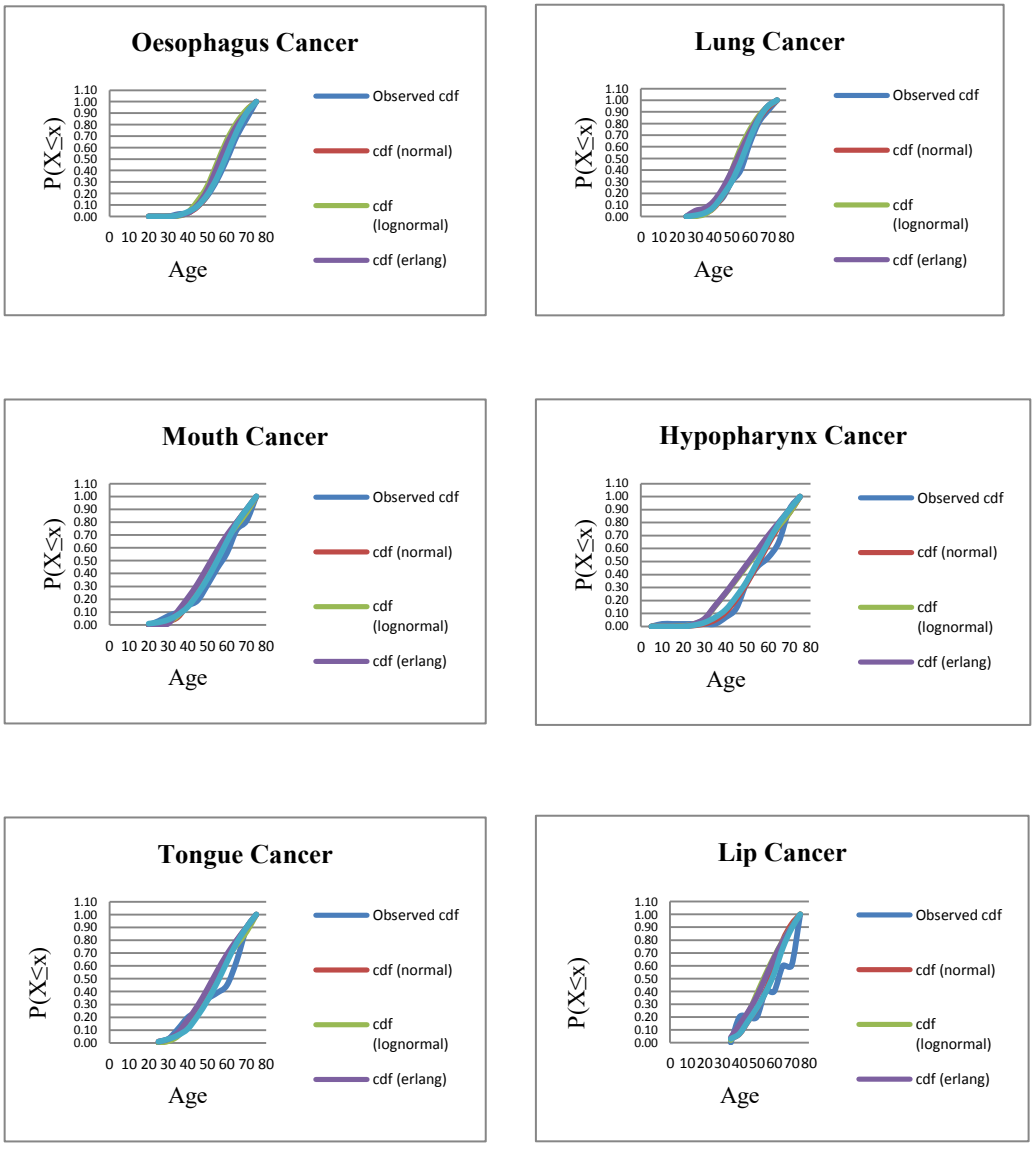
Table 3.2 (i): Fitting of probability distributions for "PHARYNX CANCER" for females (2012-14)

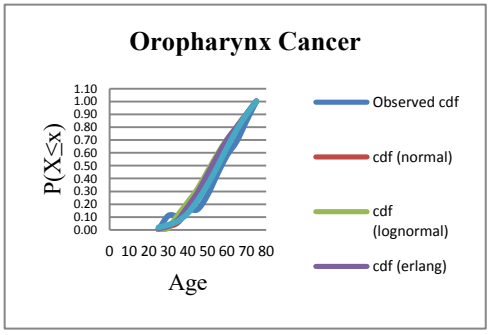
Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=50.83 σ=10.67	Lognormal μ=3.91 σ=0.20	Erlang k=27.2320 λ=0.0199	Weibull α=54.3799 β=5.0184
Below 35	0	0	0	0	1
35-40	1	1	1	1	1
40-45	0	1	1	1	1
45-50	3	1	1	1	1
50-55	1	1	1	1	1
55-60	0	1	1	1	1
60-65	0	1	0	0	1
65-70	0	0	0	0	0
70-75	1	0	0	0	0
Kolmogorov Smirnov D Statistics		0.26	0.20	0.1514	0.1889

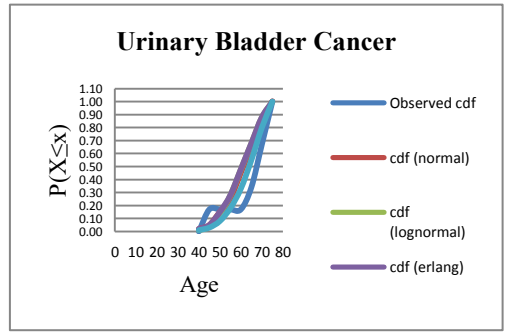
Table 3.2 (j): Fitting of probability distributions for "LARYNX CANCER" for females (2012-14)

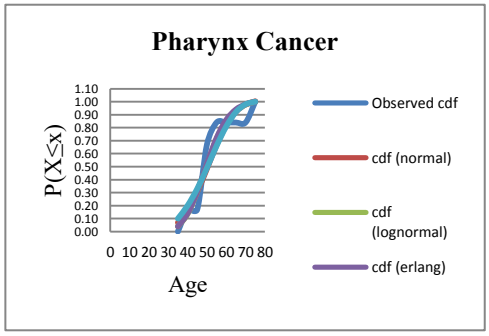
Age Interval	Observed Frequencies	Theoretical Frequencies			
		Normal μ=57.17 σ=6.94	Lognormal μ=4.04 σ=0.13	Erlang k=60.7443 λ= 0.0176	Weibull α=59.7035 β=9.5518
Below 40	0	0	0	0	0
40-45	1	0	0	1	1
45-50	2	2	2	2	2
50-55	1	3	4	4	3
55-60	6	5	4	4	4
60-65	3	3	3	3	4
65-70	2	1	2	2	2
70-75	0	0	1	0	0
Kolmogorov Smirnov D Statistics		0.10	0.14	0.1552	0.0966

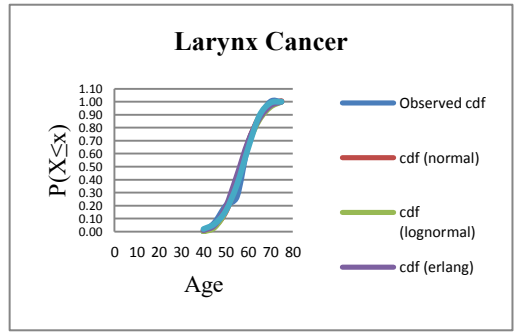
Fig.3.2 Curve for fitted and observed probability distribution functions for females (2012-14)











4. Results and Discussion:

The salient features that have been revealed from the goodness of fit tests and graphs are as follows:

In case of males, the Weibull and the Erlang Distribution provides a good fit for Oesophagus Cancer (Table 3.1(a)) and Lung Cancer data (Table 3.1(c)). The Weibull, Erlang and Normal Distribution provides a good fit for Hypopharynx Cancer (Table 3.1(b)), Tongue Cancer (Table 3.1(e)) and Urinary Bladder Cancer data (Table 3.1(i)). Only the Weibull Distribution provides a good fit for Mouth Cancer data (Table 3.1(d)). All the four probability distributions provides a good fit for Larynx Cancer (Table 3.1(f)), Oropharynx Cancer (Table 3.1(g)), Lip Cancer (Table 3.1(h)) and Pharynx Cancer data (Table 3.1(j)).

In case of females, the Weibull, Erlang and Normal Distribution provides a good fit for Oesophagus Cancer (Table 3.2(a)), Lung Cancer (Table 3.2(b)), Hypopharynx Cancer (Table 3.2(d)) and Tongue Cancer data (Table 3.2(e)). All the four probability distributions provides a good fit for Mouth Cancer (Table 3.2(c)), Lip Cancer (Table 3.2(f)), Oropharynx Cancer (Table 3.2(g)), Urinary Bladder Cancer (Table 3.2(h)), Pharynx Cancer (Table 3.2(i)) and Larynx Cancer data (Table 3.2(j)).

In general, the Log Normal Distribution does not provide a best fit of Tobacco Related Cancer as evidenced by the Kolmogorov Smirnov Test and the graph for cdf.

The Normal Distribution, although accepted to be well fitted on the basis of Kolmogorov Smirnov Test, does not seem to compete with Erlang and Weibull Distribution.

The Weibull and the Erlang cdf are quite close to the observed cdf plot. Although these two distributions are observed to be competing with each other, the Weibull Distribution happens to be the best fit for the probability distributions as evidenced by the Kolmogorov Smirnov Test and the graph for cdf.

References:

[1] Anand K., Gupta V. and Yadav K. (2010). Patterns of tobacco use across rural, urban, and urban-slum populations in a North Indian community. Indian J Community Med, 35, 245-251.

[2] Barman D., Kalita M., Kataki A.C., Sharma A. and Sharma J. D. (2016). Cancer statistics in Kamrup urban district: Incidence and mortality in 2007–2011. Indian J Cancer 53, 600-606.

[3] Bonu S. Jamjoum L. Jha P., Nguyen S. and Rani M. (2003). Tobacco use in India: prevalence and predictors of smoking and chewing in a national cross sectional household survey,Tob. Control, 12 (2003), e4, doi: 10.1136/tc.12.4. e4.

[4] Dobe M., Rahman K. and Sinha D. N. (2006). Smokeless tobacco use and its implications in WHO South East Asia Region, Indian J Public Health, 50, 70-75.

[5] Elizabeth J., Rooban T., Umadevi K. R. et al. (2010). Sociodemographic correlates of male chewable smokeless tobacco users in India: a preliminary report of analysis of National Family Health Survey, 2005-2006, Indian J Cancer, 47, 91-100.

[6] Gupta B. (2013), Burden of Smoked and Smokeless Tobacco Consumption in India - Results from the Global adult Tobacco Survey India (GATS-India), 2009-2010, Asian Pacific Journal of Cancer Prevention, 14, 3323-3329.

[7] Gupta P.C. and Ray C.S. (2003). Smokeless tobacco and health in India and South Asia, Respirology, 8, 419-431.

[8] Jindal et al.(2006).Tobacco smoking in India: prevalence, quit-rates and respiratory morbidity, Indian J Chest Dis Allied Sci., 48, 37-42.

[9] John S. R. (2005), Cancer and the family: An integrative model, Cancer, 104, 2584-2595.

[10] Kathirvel S., Sharma S. and Thakur J. S. (2014). Women and tobacco: A cross sectional study from North India, Indian Journal of Cancer, 51, S78-S82.

[11] Keeping E. S. (1962). Introduction to Statistical Inference, D. Van Nostrand Co., Inc.,Princeton, N.J.

[12] Mathur P. and Shah B. (2011). Research priorities for prevention and control of non-communicable diseases in India, Indian J. Community Med., 36, S1:72-7.

[13] Pal S. K. (1998). Statistics for Geoscientists: Techniques and Applications, Concept Publishing Company, New Delhi.