JRIIE Journal of Research Innovation and Implications in Education
Center for Research Implications & Practice

Website: www.jriiejournal.com **SSN 2520-7504 (Online) Vol.5, Iss.3, 2021 (pp. 1 - 14)**

The Impact of Smoking Risk Determinants on Lung Cancer Disease

Joie Lea Murorunkwere* & Jean Paul Nizeyimana

University of Tourism, Technology and Business Studies, Rwanda

Corresponding Author*: isimbilea@gmail.com

Abstract: *Lung cancer (LC), the leading cause of cancer-related deaths worldwide, is a complex and highly heterogeneous disease. Additional to its biological complexity, LC patients are often confronted with a high degree of stigma, mostly from the association of the disease with tobacco. Nonetheless, a proportion of LC patients are non-smokers. This paper aims to investigate the prevalence of smoking and the factors associated with it to the risk determinants of Lung Cancer, in Butaro Cancer Center. In Rwanda, this is most acutely seen in cancer, with more than a thousand registered cases of cancer between 2012 and 2020. To respond to this growing need, the first comprehensive cancer center in Rwanda has been developed, adjacent to the Butaro District Hospital. Data have been collected through questionnaires, interviews and document review. The population of 110 was stratified by social demographic types within the hospital. Multistage cluster sampling was used to draw 52 respondents. Collected data have been analysed using SPSS. Participants with smoking status had higher rates of lung cancer risks at 35%. In addition, those with less education and a family history of lung cancer and who were current smokers had higher lung cancer risk scores at 45%. Predictors of perception of synergistic risk were marital status and health-related self-concept at 20%. Based on the findings, we recommend that Lung cancer risk reduction interventions with vulnerable populations are needed. This review comprehensively assesses the current knowledge in terms of risk factors and disease characteristics in the non-smoker lung cancer population.*

Keywords*: Lung Cancer, Non-smoker, Prevention and control, Tobacco, Tuberculosis, Perceived risk*

How to cite this work (APA):

Murorunkwere, J. L. & Nizeyimana, J. P. (2021). The impact of smoking risk determinants on lung cancer disease. *Journal of Innovation and Implications in Education, 5(*3), 1 -14.

1. Introduction

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body (NIH, 2021). Lung cancer is a type of cancer that begins in the lungs. Smoke is a visible vapor from a burning substance and cigarette smoke is a complex mixture of chemicals produced by burning tobacco and the additives. The smoke contains compounds of a different physicochemical nature and the degree of harmfulness of some of these compounds are very dangerous since they can cause lung diseases and heart disease (Brook, et al.; 2014). Researchers have estimated that cigarette smoke has 7,357 chemical compounds from many different classes (Rodgman and Perfetti, 2009). In Rwanda, this is most acutely seen in cancer, with more than three thousand registered cases of cancer between 2017 and 2020.Currently, more than 1700 young and old patients are being treated by Doctors at the Partners in Health-supported facility every year at Butaro District Hospital. Tobacco consumption is low in Rwanda and there are few tobacco-related tumours, such as lung and laryngeal

cancer. Other tumours believed to be associated with aspects of Western life-style, such as colorectal and breast cancer, are also relatively infrequent. To respond to this growing need, Partners in Health and the Ministry of Health developed the first comprehensive cancer center in East Africa, adjacent to the Butaro District Hospital. Every year, smoking accounts for more than 7 million preventable deaths worldwide (World Health Organization, 2019).

This study was guided by the following research questions:

a) What are the factors associated with smoking risk determinants of Lung Cancer?

b) What is the relationship between smoking determinants and lung cancer risks?

2. Literature Review

The annual deaths are expected to reach 8 million by 2030 if no cost effectiveness measures to reduce smoking are initiated (Blecher & Ross, 2018). Approximately, 80% of all the tobacco attributable deaths occur in low-middle income countries (LMICs) (Fontham,2014) such as Rwanda where tobacco use among adults is estimated to be 13% (Ministry of Health Rwanda 2015). In South Africa 2018, smoking was a causal agent in the distribution of mortality (Gruenewald, et al.; 2017) and in terms of the overall deaths in South Africa, Groenewald, et al. (2007) estimated that about 8.5% of all deaths in 2000 can be attributed to smoking. Furthermore Sitas et al. (2014) stated that smoking ranked third in terms of mortality among 17 risk factors evaluated, in 2018 about 12% of adult (greater than 25 years) died because of smoking.

The environmental tobacco smoke, also known as passive smoke or secondhand smoke contains the same harmful chemicals as the smoke that smokers inhale, occurs when nonsmokers inhale other people's tobacco smoke. There is strong evidence that environmental tobacco smoke causes serious damage to human health. Several epidemiological investigations (see, for example, Fonthan et al. 2014; Bennett et al, 2016; Malats et al, 2018; Miller et al, 2017; Environmental Protection Agency, 2016; and National Cancer Institute, 2016) have demonstrated that environmental tobacco smoke contributes to the following health effects: carcinogenic (lung and nasal sinus cancer), cardiovascular (heart disease mortality, acute and chronic coronary heart disease morbidity), respiratory (in children: acute lower respiratory tract infections, asthma induction and exacerbation, chronic respiratory symptoms, middle ear infections; in adults: eye and nasal irritation), and developmental (foetal growth: low birth weight or small for gestational age, sudden infant death syndrome).

The National Cancer Institute (2016), also states that the effect associated with environmental tobacco smoking exposure include cervical cancer, exacerbation of cystic fibrosis, decreased pulmonary function, spontaneous abortion and an adverse impact on cognition and behavior.

2.1 Correlation between Smoking and Lung Cancer

Lung cancer evolves as a result of a series of mutational events that have been studied in detail by numerous investigators (Sato et al.,2017). However, the molecular pathogenesis of lung cancer remains incompletely defined. Because inflammation appears to play an important role in the pathogenesis of lung cancer, a thorough understanding of lung cancer pathogenesis requires consideration of the tumor micro-environment (TME) and the inflammatory pathways operative in carcinogenesis (Prendergast GC., 2018). The tobacco-induced pulmonary cellular network presents a unique environment in which carcinogenesis proceeds in complicity with surrounding lung inflammatory, structural, and stromal cells. The pulmonary diseases that are associated with the greatest risk for lung cancer are characterized by abundant and deregulated inflammation (Sevenoaks and Stockley, 2016). The commonalities in smoking and lung cancer begin with the profound alterations induced by cigarette smoke, which contains known carcinogens as well as high levels of reactive oxygen species (ROS). The ready induction of reactive oxygen species after tobacco smoke exposure leads to impairment of epithelial and endothelial cell function as well as inflammation (Woods, et al, 2018).Inhalation transports tobacco-specific carcinogens more distally toward the bronchoalveolar junction where adenocarcinoma often arises. Secondly, blended reconstituted tobacco releases a higher concentration of N-nitrosamines from tobacco stems (Calle, et al., 2017). A relatively older estimate of more than 26,000 cases from 17 published reports suggests that the adenocarcinoma to squamous-cell carcinoma ratio is approximately 0.4 lung cancers in smokers as compared to 3.4 in never-smokers (Sun S et al., 2017). Lung cancer risk increases with the duration and intensity of tobacco consumption (Cancer Research UK., 2018).

2.2. Patient Characteristics, Environmental Factors, and Lung Cancer

Certain patient characteristics have consistently shown an impact on lung cancer outcomes. For example, lung cancer is a disease of the elderly, although advancing age was not a prognostic factor for survival but high scores on the Charlson Comorbidity Index (CCI) were a factor. Taken together, toxicity, age and high CCI scores were significant predictors (Gregory, et al.,2016). The

incidence of lung cancer is higher among men (34%) as compared to women (13.5%). The age-standardized ratio for cancer incidence is 33.81%, and for mortality is 29.2% in men alone (Forman, et al., 2017).

In the past, the incidence was lower in females, but worldwide it is now the fourth most frequent cancer in women (516,000 cases; 8.5% of all cancers) and the second most common cause of cancer deaths (427,000 deaths; 12.8% of the total) (*GLOBOCAN 2008).* The highest incidence rate in women is observed in North America, where lung cancer is now the second most frequent cancer in women. This is attributed to smoking. It is the lowest in central Africa, where it is the 15th most frequent cancer in women. As one in 5 women who develop lung cancer is a never-smoker, it remains a mystery as to what exactly causes their cancer.

Lung cancer in never-smokers is proposed to be due to multiple risk factors, including genetic predisposition—although this is exceedingly rare (1% with >3 affected relatives). Genetics mutations remain an underlying cause as we do encounter lung cancer at a relatively earlier age when it runs in families. Among the first studies revealing a genetic link was one conducted over 10 years ago by Tokuhata *et al*., 2009. The study revealed that never-smokers with lung cancer were 40% more likely than never-smoking controls to report a first degree relative with lung cancer. Women were more likely to report such a family history and 10–15% had at least one first-degree relative with the disease

3. Methodology

3.1 Research Design

This study adopted concurrent mixed methods research design. The study employed concurrently qualitative and quantitative methods. In this context the concurrent mixed approach is used according to Gall and Borg (2010).

3.2 Population and Sampling Techniques

The data surveyed was carried out in Butaro Cancer Center of Excellence in Burera District, and the population of 110 was stratified by social demographic types within the hospital. Multistage cluster sampling with the probability proportional to size was used to draw 52 respondents.

To determine the sample size, the study adopted the formula of Alain Bouchard as follows:

nc= n /(1+n /N)

Where nc = The sample to be determined; N= the number of whole population which is 110; n=the target of population which is 96.

We have nc= 96/(1+96/110) = 51.52 thus the sample size was 52

3.3 Research Instruments

In order to collect data, the study used questionnaire, interview and documents analysis. Questionnaire has been administrated to the patients and their caregivers and interview has been conducted between the researchers, hospital Doctors and Nurses. Document's analysis was used of outside sources, internet, written documents, to support the viewpoint or argument on Smoking Risk Determinants on Lung Cancer Disease. The collected data have been coded and analyzed using the Statistical Package for the Social Sciences (SPSS).

3.4 Validity and Reliability

To ensure the validity of research instruments, the questionnaire has been designed in line with the research objectives. The questionnaire has been pretest to identify and remove any ambiguous questions. Concerning reliability, research instruments have been pretest by 10 respondents randomly selected from the sample. During this exercises offensive and ambiguity detected in any questions was adjusted.

4. Results and Discussion

4.1 Description of Data

Rwandan young adults smoking status, perceptions of the health effects of nicotine and cigarettes, attitudes toward smoking control and awareness of Government health warning on the harmful effects of smoking were surveyed by means of a series of interviewer-administered questionnaires conducted by fieldworkers. The study shows that 8% of Rwandan youth are smokers compared to 13% of the general population (15–64 years) (National Institute of statistics. Rwanda Demographic and Health Survey. Rwanda. 2015. 640 p.). This relatively high prevalence among the youth raises concerns because of the negative long-term negative effects of tobacco use including death. The tobacco industry has continued to aggressively use cutting-edge technology to market their products and recruit more users among youth. Additionally, young people may have strong social networks which could influence initiation of tobacco consumption while making it difficult for those who have started to quit smoking.

Therefore, behavioral interventions coupled with cessation programs would be an important step to support these young people to avoid or quit smoking (Roberts, et Al. 2018) The data surveyed was carried out in Butaro Cancer Center of Excellence in Burera District, and the population of 110 was stratified by social demographic types within the hospital. Multistage cluster sampling with the probability proportional to size was used to draw 52

respondents. Participants with smoking status had higher rates of lung cancer risks at 35%. In addition, those with less education and a family history of lung cancer and who were current smokers had higher lung cancer risk scores at 45%. The study variables are gender, age, social status, marital status, smoking status and education level.

The 52 clusters were considered to be a random selection of clusters from Butaro Cancer Center of Excellence in Burera District. Then the random selection of respondents was drawn from the clusters, for each selected respondent a sampling weight was calculated, using the stratification variables and type of area and by post-stratification for age, gender, marital status and education. The respondents in the survey were asked to recall a number of anti-smoking messages which appeared as warning messages on cigarette advertisements. There were different warning messages appearing on cigarette packets, television, radios, magazines and at hospital posters:

Danger: Smoking can kill you, Smoking causes cancer and Smoking Damages Your Lungs.

Warning: Don`t smoke near children, your smoking can harm those around you.

Pregnant? Breast Feeding? Your Smoking Can Harm Your Baby

Danger: Smoking Causes Heart Disease and Tobacco is addictive and causes cancer.

The main aim of the study was to investigate the occurrence of smoking and the factors associated with it and to determine the risk factors associated with Lung Cancer. Several risk factors have been linked to LC in non-smokers. For each respondent the total number of messages spontaneously recalled was noted. This is the dependent variable after taking the record of it to overcome the independent variables in the data. We are interested in applying the various techniques to estimate the variance components by fitting variance component models to the data and comparing them to assess their advantages and disadvantages. As a result, to determine the effects of various explanatory variables on the number of messages recalled, we make use of ANOVA, the linear mixed model and the linear model.

4.1.1 Box-and-whisker-plots

We now consider the box plots as a graphical representation of the data to check for outliers, trends and differences between groups of the same variable, for example differences in the education levels with respect to the dependent variable (logresp) can give us insight and summary between the groups of individuals with different educational levels. The dependent variable is plotted on the y-axis whilst the other variables like gender, age, marital status, smoking status and educational status will be on the x-axis. Let us now look at these plots done by SPSS:

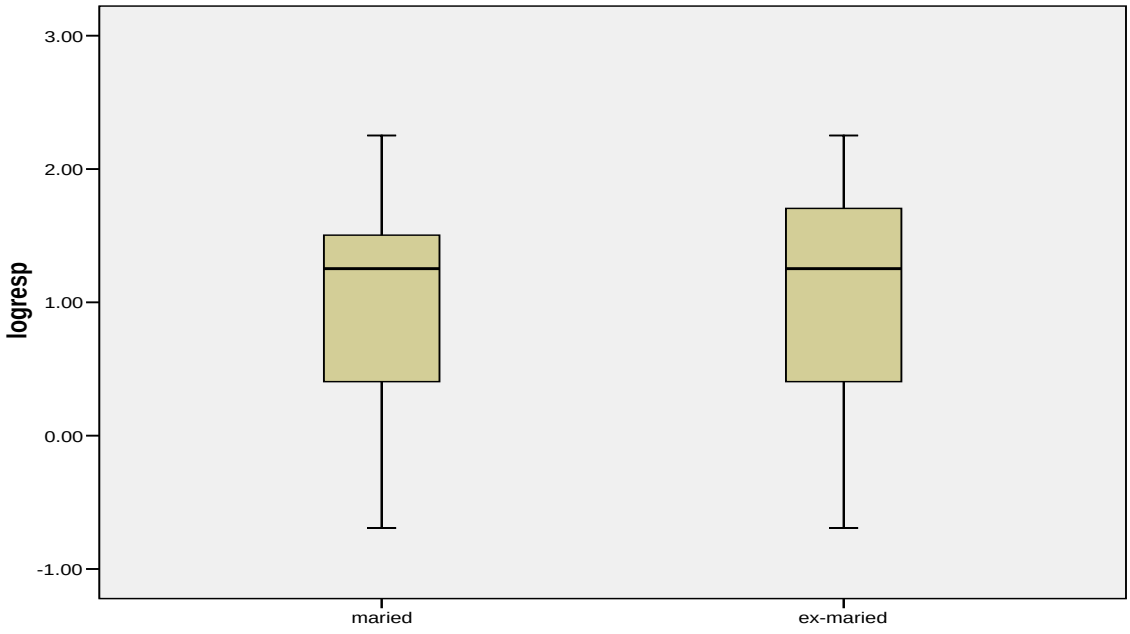

Figure 1. Graph of marital status vs log resp

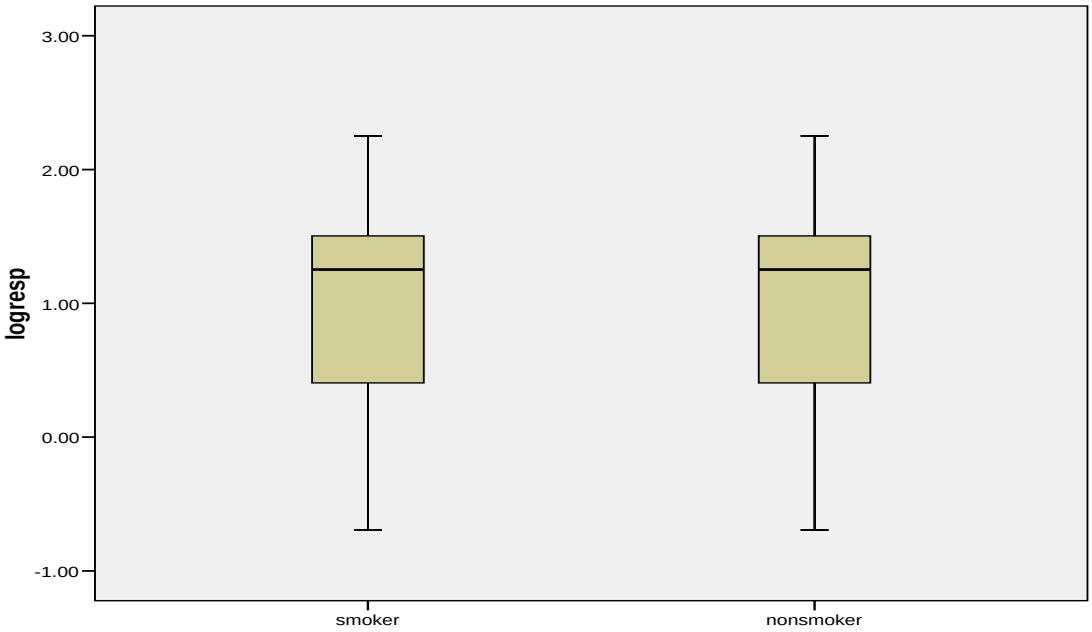

Figure 2. Graph of smoking status vs log resp

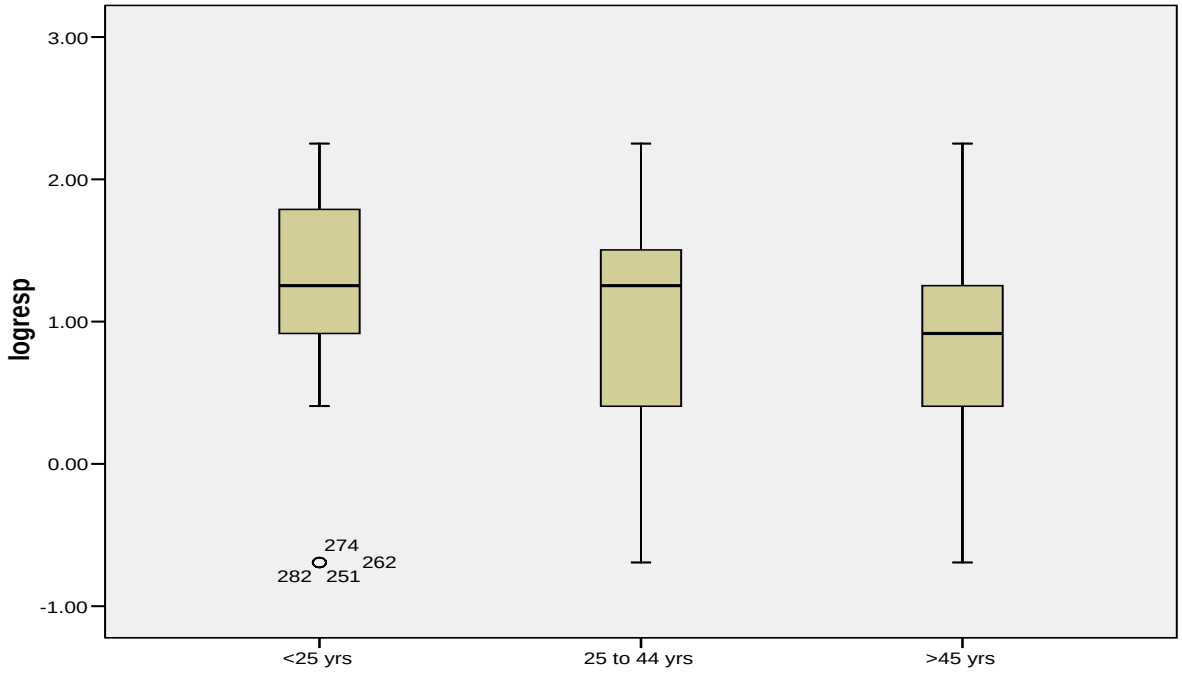

Figure 3. Graph of age vs log resp

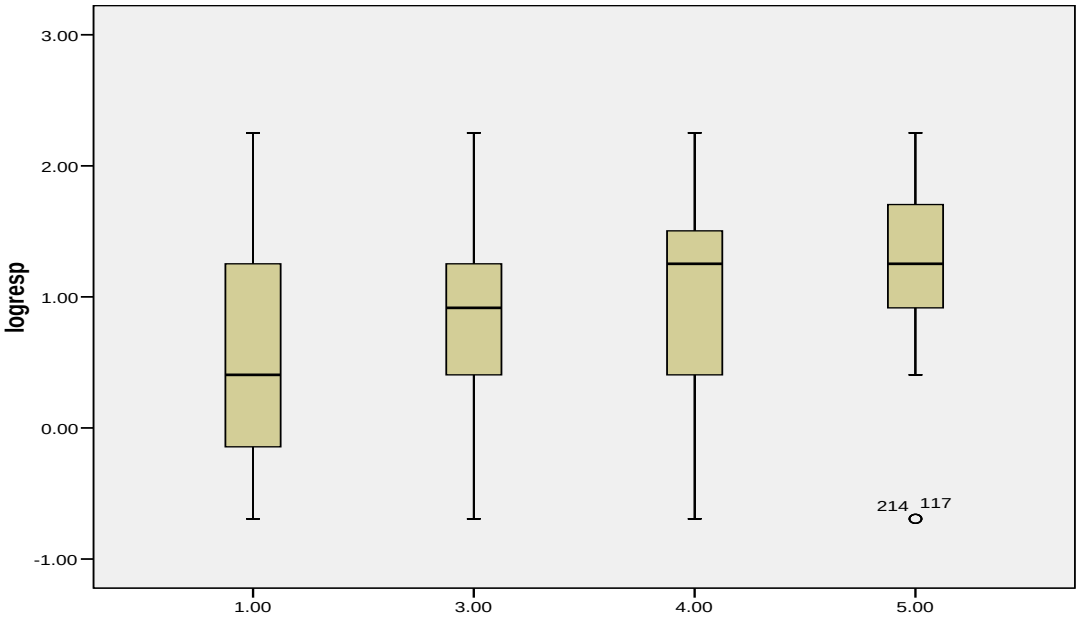

Figure 4. Graph of education status vs logresp

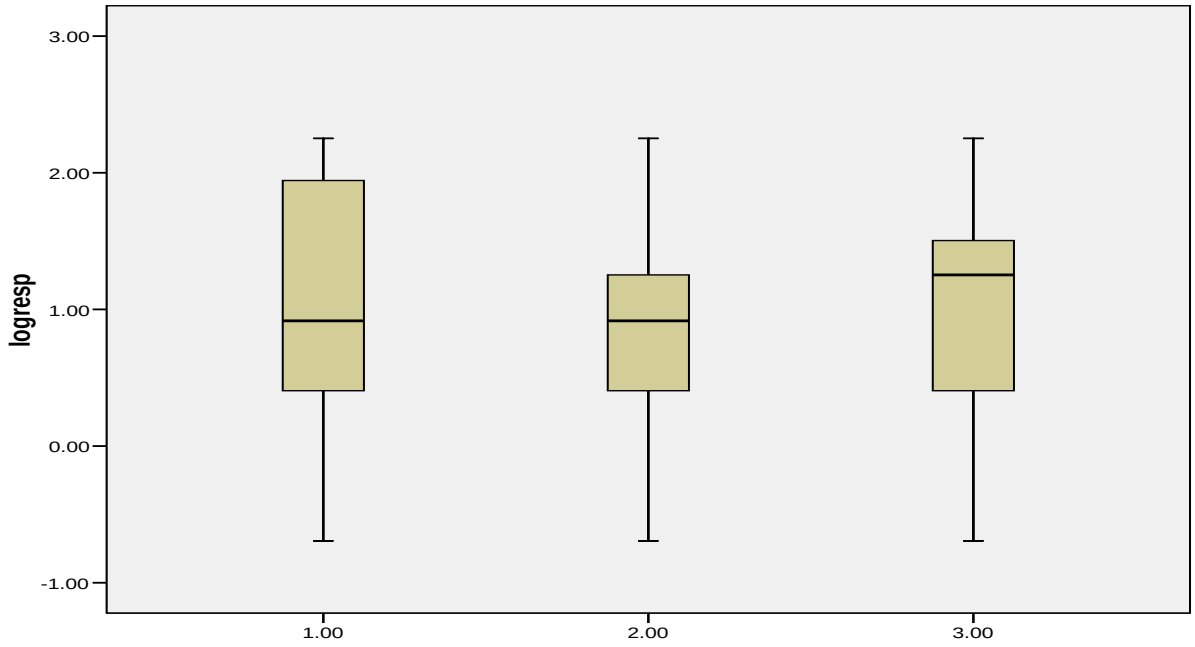

Figure 5. Graph of social vs logresp

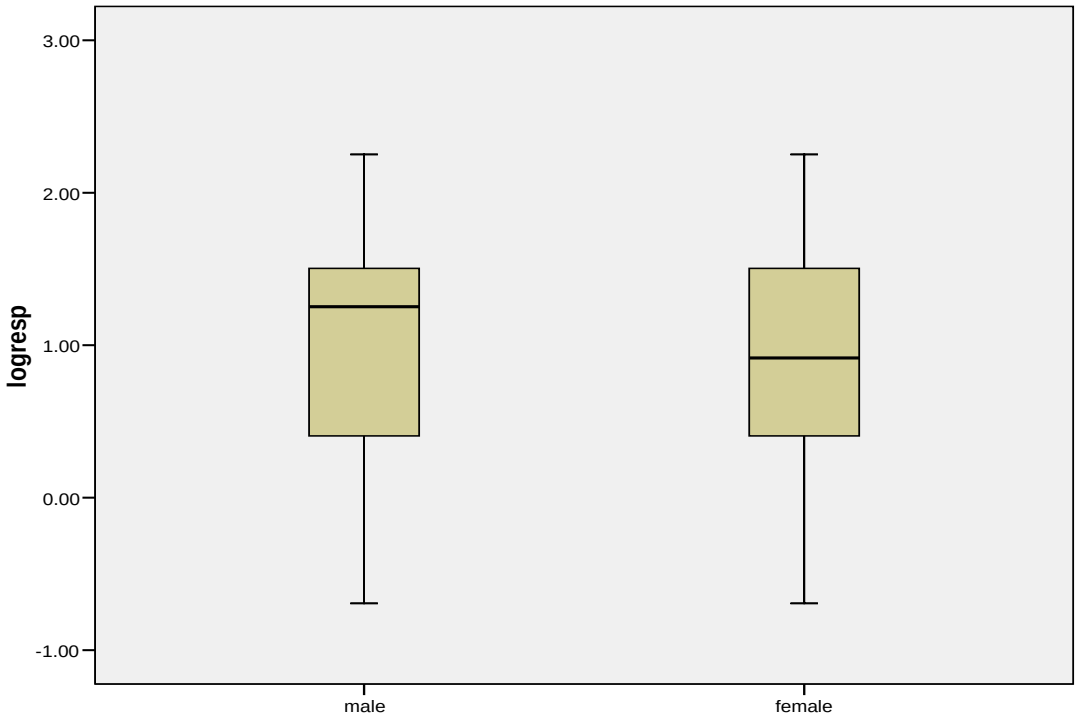

Figure 6. Graph of gender vs logresp

Figure 1 shows the two marital status groups namely the married and ex-married to have the number of messages recalled to be skewed to lower number of messages recalled and their medians are almost equal.

Figure 2 shows smokers and non-smokers recalled a low number of anti-smoking messages having their distributions skewed to a low logresp value.

Figure 3 shows us that the age group from 25-44(level 2) years are skewed to lower number of messages recalled whilst the < 25 (level 1) year's group are skewed toward a higher number of messages recalled. The >45 (level 3) years group are about symmetrical in their distribution. The 25-44 (level 2) age group recalled the highest number of messages while the >45 (level 3) age group recalled the lowest number of messages.

Figure 4 shows that the highly educated respondents recalled more anti-smoking messages than the lesser educated people. The 4th group has distribution that is skewed toward a lower number of messages recalled whilst the 1st group has distribution that is skewed toward a higher number of messages recalled and the 2nd group is symmetric in distribution. Also, the median number of messages recalled increased as the level of education increased.

Figure 5 shows that the 1st and 3rd groups have distributions that are skewed towards a lower number of messages recalled whilst the 2nd group is symmetric in distribution. The 3rd group has the higher median number of messages recalled and the 1st & 2nd groups have the same median number of messages recalled.

Figure 6. This graph clearly shows us that the distribution of the males and females are very similar whilst only the medians differ slightly. The median number of messages recalled was higher for males than for females. The distribution of the females seems to be symmetrical in the number of messages that they recalled.

4.2 Smoking and lung cancer Interaction

Smoking tobacco is the number one risk factor associated with lung cancer (CDC, 2020). Lung cancer develops when cells in the lungs become damaged and grow uncontrollably, causing tumors that can make it difficult to breathe. In this study, we looked at the risk of smokers developing lung cancer. The researchers found that the average risk for nonsmokers over the study period was 1.8% for males and 1.3% for females. The risk jumped to 14.8% for males and 11.2% for females in current smokers.

4. 3 Generalized Linear Model

The Generalized Linear Models are a family for important models for categorical responses as well as standard models for continuous responses. The generalized linear model can be seen as an extension of linear multiple regression for a single dependent variable, and understanding the multiple regression is fundamental to understanding the general linear model (Diggle et al. 2002). The Generalized Linear Models, according to

McCullagh and Nelder (1989) are one such family of models and are generally suitable for discrete repeated measurements in the context of correlated data.

4.3.1 The Exponential Family

In generalized linear models, the response is assumed to possess a probability distribution of the exponential form. That is, the probability density of the response *Y* for continuous response variables, or the probability function for discrete responses, can be expressed as

$$f(y_i, \theta_i, \phi) = exp\left\{\frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi)\right\} \quad (4.1)$$

for some functions *a*, *b*, and *c* that determine the specific distribution. For fixed ϕ, this is a one parameter exponential family of distributions. The functions *a* and *c* are such that $a(\phi) = \phi/w_i$ and $c = c(y_i, \frac{\phi}{w_i})$, where w_i is a known weight for each observation.
Standard theory for this type of distribution gives expressions for the mean and variance of *Y*.

$$E(Y_i) = b'(\theta_i) \qquad Var(Y_i) = \frac{b''(\theta_i)\phi}{w_i}$$

where the primes denote derivatives with respect to θ_i. If μ represents the mean of *Y*, then the variance expressed as a function of the mean is

$$Var(Y_i) = \frac{V(\mu)\phi}{w_i}$$

where *V* is the *variance function*.

4.3.2 Binomial random variable

For a binomial variable Y denoting the number of successes in n independent trials with a probability of success p in each trial, the probability distribution is:

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y} = \exp\left[ylog\frac{p}{1-p} + nlog(1-p) + log\binom{n}{y}\right]$$

From the Eq (4.1) it follows that the corresponding canonical (natural) parameter is

$\theta_i = \log\frac{p}{1-p}$ also known as the $logit(p)$. Alternatively, $\theta_i = log\frac{\mu}{n-\mu}$ where $\mu = np$

Note that in terms of θ_i

$$p = \frac{\exp(\theta_i)}{1+\exp(\theta_i)}$$

And

$$1 - p = \frac{1}{1+\exp(\theta_i)}$$

In terms of the structure of an exponential family probability density function (p.d.f) $\psi(\theta_i) = -nlog(1-p) = nlog(1+\exp(\theta_i)), \phi = 1.$ And $c(y, \phi) = log\binom{n}{y}$

Furthermore

$$E[Y_i] = \psi'(\theta_i) = n\frac{\exp(\theta_i)}{1+\exp(\theta_i)} = np$$

And

$Var(Y_i) = \psi''(\theta_i)\phi = \frac{nexp(\theta_i)(1+\exp(\theta_i)) - nexp(\theta_i)\exp(\theta_i)}{(1+\exp(\theta_i))^2}$
$= np(1-p)$. Thus, in this case
$v(\mu) = \mu\left(1 - \frac{\mu}{n}\right)$ since $\mu = np$

The Bernoulli (random variable) model for the binary response is a special case of the Binomial random variable with $n = 1$ and therefore both share the same canonical or natural parameter (McCullagh and Nelder, 1989; Molenberghs and Verbeke, 2005).

4.3.3 Examples using Generalized linear models

Logistic and Probity regression for Binary data

The natural link is the logit link so that if $Y_i \sim Bernoulli(\pi_i)$ then the linear model is

$$\ln\left[\frac{\pi_i}{1-\pi_i}\right] = x_i'\beta$$

where in terms of covariates

$$\pi_i = \frac{\exp(x_i'\beta)}{[1+\exp(x_i'\beta)]}$$

Note that the natural parameter is a function of the covariate x_i. Alternatively for the probit link, one uses the model $\Phi^{-1}(\pi_i) = x_i'\beta$ so that $\pi_i = \Phi(x_i\beta)$ where Φ denotes the distribution function of a standard normal random variable. For a Binomial variable the $Y_i \sim B(n_i, p_i)$ and the regression model is of the form logit $(p_i) = x_i'\beta$.

4.3.4 Poisson Regression for counts

The logarithm is the natural link function, leading to the classical Poisson regression model $Y_i \sim Poisson(\mu_i)$ with $\ln(\mu_i) = x_i'\beta$ where μ_i is the mean occurrence rate. This also implies $\mu_i = \exp(x_i'\beta)$ is a quantity which is always non-negative.

4.3.5 Mean and Variance Functions for the Random Components

General expressions for $E(Y_i)$ and $Var(Y_i)$ use terms in (4.1). Let $L_i = logf(y)$ denote the contribution of y_i to

the log likelihood; that is, the log-likelihood function is $L = \sum_i L_i$. Then, from (4.1),

$$L_i = \frac{[y_i\theta_i - b(\theta_i)]}{a(\phi)} + c(y_i, \phi) \quad (4.2)$$

Therefore,

$$\frac{\partial L_i}{\partial \theta_i} = \frac{[y_i - b'(\theta_i)]}{a(\phi)} \quad , \quad \frac{\partial^2 L_i}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)},$$

Where $b'(\theta_i)$ and $b''(\theta_i)$ denote the first two derivatives of $b(.)$ evaluated at θ_i. We now apply the general likelihood results

$$E(\partial L/\partial\theta) = 0 \quad \text{and} \quad -E(\partial^2 L/\partial\theta^2) = E(\partial L/\partial\theta)^2$$

Which hold under regularity conditions satisfied by the exponential family (Cox and Hinkley, 1974). From the first formula applied with a single observation, $\frac{E[Y_i - b'(\theta_i)]}{a(\phi)} = 0$, or

$$\mu_i = E(Y_i) = b'(\theta_i). \quad (4.3)$$

From the second formula,

$$\frac{b'(\theta_i)}{a(\phi)} = E\left[\left(Y_i - b'(\theta_i)\right)/a(\phi)\right]^2 = Var(Y_i)/[a(\phi)]^2,$$

So that

$$Var(Y_i) = b''(\theta_i)a(\phi).$$

In summary, the function $b(.)$ in (4.1) determines moments ofY_i.

Application

The fitted model has its explanatory variables of gender, age, social status, marital status, smoking status and educational status. The socio-demographic variables categories are encoded as gender (Male and Female); age (<25 yrs, 25 to 44 yrs and >45 yrs); Marital status (married or ex-married); Smoking status (smoker or nonsmoker); and education level on a four-point scale (primary, secondary, higher secondary and university). It should also be stated that ex-smokers were categorized as nonsmoker in this study as the smoking variable only involves the two categories. In other words, the study only concentrates on the current smoking status.

Table 1: The fixed effects can be summarized into the following table

FIXED EFFECT	SURVEY CODE	ABBREVIATION
gender (2 levels)	1=male 2=female	gen
Age (3 levels)	1:<25yrs 2:25-44yrs 3:>44yrs	age
social status (3 levels)	1=lowest …. 3=highest	Lsm
marital status (2 levels)	1=married 2=ex-married	mar
smoking status (2 levels)	0=smoker 1= nonsmoker	Ys
education on a 4-point scale (4 levels)	1=lowest … 4=highest	Edu

By default, PROC GENMOD uses a corner point parameterization for categorical variables where the last category of each variable is used as the reference category.

The output from SAS using PROC GENMOD result in the following table:

Table 2: Analysis of Parameter Estimates using Proc Genmod.

Analysis Of Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept		1	1.2468	0.1858	0.8826	1.6111	45.01	<.0001
Male	**1**	1	0.0291	0.0873	-0.1419	0.2002	0.11	0.7386
Ref=Female	**2**	0	0.0000	0.0000	0.0000	0.0000	.	.
<25 yrs	**1**	1	0.2229	0.1724	-0.1151	0.5608	1.67	0.1961
25-44 yrs	**2**	1	0.0647	0.1205	-0.1715	0.3009	0.29	0.5913
Ref= >44 yrs	**3**	0	0.0000	0.0000	0.0000	0.0000	.	.
Lowest	**1**	1	-0.0835	0.1656	-0.4080	0.2411	0.25	0.6142
Medium	**2**	1	-0.2105	0.1337	-0.4725	0.0515	2.48	0.1154
Ref=Highest	**3**	0	0.0000	0.0000	0.0000	0.0000	.	.
Married	**1**	1	-0.1479	0.1003	-0.3445	0.0487	2.17	0.1404
Ref=Ex-married	**2**	0	0.0000	0.0000	0.0000	0.0000	.	.
Smoker	**0**	1	-0.0939	0.0934	-0.2771	0.0892	1.01	0.3147
Ref=Nonsmoker	**1**	0	0.0000	0.0000	0.0000	0.0000	.	.
Lowest	**1**	1	-0.5585	0.1834	-0.9180	-0.1990	9.27	0.0023*
2nd level	**2**	1	-0.4252	0.1339	-0.6877	-0.1628	10.08	0.0015*
3rd level	**3**	1	-0.1943	0.1137	-0.4172	0.0285	2.92	0.0875
Ref=Highest	**4**	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.7590	0.0290	0.7042	0.8179		

The above table shows that the intercept for the model fitted was significant with the p-value<0.0001. Males versus Females are different with respect to the log (resp) of the number of messages recalled at the 5% level. The table further shows that people less than 25 years and those between age of 25-44 years versus peoples with age greater than 44 years are not different with respect to log of the number of messages recalled at the 5% level. New social 1&2 versus new social 3 groups are not different with respect to the log of the number of messages recalled at the 5% level. Married people versus ex-married people were not different with respect to log of the number of messages recalled at the 5% level. Smokers versus non-smokers were not different with respect to log of the number of messages recalled at the 5% level of significant. People with lowest level and those with 2nd level of education versus people with highest level of education are different with respect to the log of the number of

messages recalled at the 5% level. There were no other differences in new education group with respect to the log (resp).

4.3.6 Cluster Sampling Methods

Cluster sampling is a sampling technique in which the entire population of interest is divided into groups, or clusters, and a random sample of these clusters is selected. Each cluster must be mutually exclusive and together the clusters must include the entire population. After clusters are selected, then all units within the clusters are selected. No units from non-selected clusters are included in the sample. This differs from stratified sampling, in which some units are selected from each group. When all the units within a cluster are selected, the technique is referred to as **one-stage cluster sampling**. If a subset of units is selected randomly from each selected cluster, it is called **two-stage cluster sampling**. Cluster sampling can also be made in three or more stages: it is then referred to as multistage cluster sampling. The main reason for using cluster sampling is that it usually much cheaper and more convenient to sample the population in clusters rather than randomly. In some cases, constructing a sampling frame that identifies every population element is too expensive or impossible. Cluster sampling can also reduce cost when the population elements are scattered over a wide area.

Here we look at only the one stage cluster sampling (with unequal sized clusters) since we select all the units within the cluster. If a population consists of M clusters in sizes $N_1, N_2, \dots, N_n \quad (\sum_{i=1}^{n} N_i = N)$, then sample a simple random sample of m chosen clusters and then estimate $\overline{Y}$. There are three estimators to be considered:

(a) Cluster Sample Ratio

$$\overline{y}_{c(a)} = \frac{\sum_{i=1}^{m} y_{iT}}{\sum_{i=1}^{m} n_i}$$

where

- $\overline{y}_{c(a)}$ is biased
- $Var\left(\overline{y}_{c(a)}\right) = \frac{(M-m)M}{(M-1)n} \sum_{i=1}^{m} \left(\frac{N_i}{N}\right)^2 \left(\overline{Y}_i - \overline{Y}\right)^2$

  estimated by $s^2\left(\overline{y}_{c(a)}\right) = \frac{(M-m)M}{m(m-1)} \sum_{i=1}^{m} \left(\frac{N_i}{N}\right)^2 \left(\overline{y}_i - \overline{y}_{c(a)}\right)^2$

(b) Cluster Sample Total (used to estimate the mean not the total)

$$\overline{y}_{c(b)} = \frac{M}{Nm} \sum_{i=1}^{m} y_{iT}$$

where

- $\overline{y}_{c(b)}$ is unbiased
- $var\left(\overline{y}_{c(b)}\right) = \frac{(M-m)M}{(M-1)mN^2} \sum_{i=1}^{m} \left(y_{iT} - \overline{Y}_T\right)^2$ where $\overline{Y}_T = \frac{N}{M}\overline{Y}$

(c) Unweighted Average of chosen cluster means

$$\overline{y}_{c(c)} = \frac{1}{m} \sum_{i=1}^{m} \overline{y}_i$$

- Biased and inconsistent
- $Var\left(\overline{y}_{c(c)}\right) = \frac{(M-m)}{mM(M-1)} \sum_{i=1}^{m} \left(\overline{y}_i - \overline{y}_c\right)^2$ estimated by $s^2\left(\overline{y}_{c(c)}\right) = \frac{(M-m)}{mM(M-1)} \sum_{i=1}^{m} \left(\overline{y}_i - \overline{y}_{c(c)}\right)^2$

In order to estimates the survey means, we used SAS PROC SURVEYMEANS procedure. When computing these estimates, the procedure takes into account the sample design used to select the survey sample. The sample design can be a complex survey sample design with stratification, clustering, and unequal weighting. In addition to estimates for the entire survey population, the procedure can compute estimates for population subgroups. The SURVEYMEANS procedure uses the Taylor expansion method for estimating sampling errors of estimators based on complex sample designs. This method obtains a linear approximation for the estimator and then uses the variance estimate for this approximation to estimate the variance of the estimate itself (Woodruff 1971, Fuller 1975).

Now we look at the results of the data obtained by using PROC SURVEYMEANS procedure. The reason for doing this was to estimate the means for Cluster Sampling.

The output from SAS using PROC SURVEYMEANS result in the following table:

Table 3: Analysis of Parameter Estimates of survey means

Statistics			
Variable	**Level**	**Mean**	**Std Error of Mean**
Gender	**1**	0.454810	0.030339
	2	0.545190	0.030339
New age	**1**	0.139942	0.023583
	2	0.693878	0.028170
	3	0.166181	0.023351
New social	**1**	0.128280	0.034332
	2	0.282799	0.037224
	3	0.588921	0.054255
New martial	**1**	0.626822	0.036280
	2	0.373178	0.036280
smoke	**0**	0.676385	0.030482
	1	0.323615	0.030482
New educat	**1**	0.116618	0.023598
	3	0.279883	0.032233
	4	0.364431	0.022285
	5	0.239067	0.030867

The above table shows the means of the fixed effects together with their standard errors with respect to the log of the number of messages recalled at the 5% level.

4.4 Discussion

The issue of health warning related to smoking is an ongoing campaign and the effect that it was intended for has been achieved in certain parts of the world such as Sub-Saharan and West Europe. Nevertheless, the mindsets of humans need to be constantly fashioned via effective communication and health warnings related to smoking.

The above findings show that marital status, age and gender are non-significant factors for anti-smoking awareness.

Respondents with lower level of education are more likely to say they are not aware at all about the harmful effects of smoking. This is not surprising, since almost all smoking warning messages are written messages. The messages are not like the commercial adverts which appeal to everyone`s eye. In other words, whilst everyone from illiterates to the bookworms are exposed to the dangers of smoking, the written health warning messages on the cigarette packs are overshadowed by the beauty of the cigarette trademarks.

Findings showed that smokers do not recognize the harmful effects of cigarette smoke as the non-smokers do.

This is because smokers are smoking cigarettes with more ignorance of the danger of smoking than non-smokers. Respondent's shows that lower social class and third social class recalled less anti-smoking messages while the upper social class recalled more anti-smoking messages. One possible interpretation of these results is lower social class, are smoking cigarettes with more ignorance of the danger of smoking than the upper social class smokers. Moreover, smokers from the lower social class are more likely not to be aware about the harmful effects of smoking than the non-smokers of the same social class. This show that the health warning messages are not well campaigned to lower social class. The other important result reflected in this study is the importance of education for the healthy and well-informed society.

The Generalized Linear Model showed that only peoples less than 25 years and those between ages of 25-44 years versus peoples with age greater than 44 years different with respect to the number of messages recalled at the 5% level of confidence. The fixed effects were not different with respect to the number of messages recalled at 5% level of confidence. The Linear Mixed Model showed that education was significantly different with respect to the number of messages recalled at 5% level of confidence and the other fixed effect was not different. In fact, both models got the same results with respect to education level.

5. Conclusion and Recommendations

Two approaches were used to evaluate the effect of smoking on lung cancer risk determinants. With the use of multiple logistic regression, a model was constructed of the effects of smoking, education level, duration of smoking, marital status, and smoking cessation on lung cancer risk.

This study readily confirmed the strong relationship between cigarette smoking and lung cancer. The risk of lung cancer was primarily determined by the habit of smoking cigarette, but it was modified by smoking practices as smoker and non-smoker, inhalation practices, and age. Primary smoking risks have a more consistent negative effect on lung cancer. However, the findings suggest that the risk of lung cancer in smokers can best be reduced by cessation of smoking. In addition to that, we recommend that Lung cancer risk reduction interventions with vulnerable populations are needed. This review comprehensively assesses the current knowledge in terms of risk factors and disease characteristics in the non-smoker lung cancer population.

References

Bennett W.P. (2016). Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. J Natl Cancer Inst, 91, 14-2016.

Brook R.D. (2014). Air pollution and cardiovascular disease: a statement for healthcare professionals from the Expert Panel on Population and Prevention Science of the American Heart Association. Circulation, 109, 2655-2671.

Centers for Disease Control and Prevention (2020), What Are the Risk Factors for Lung Cancer? Available at https://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm

Duchateau, L. (2018). "Linear Mixed Models: An Introduction with Applications in Veterinary research', ILRI (International Livestock Research Institute) Nairobi, Kenya.

Environmental Protection Agency (EPA) (2016). Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders. EPA/600/6-90/006F.

Fontham E.T. (2014). Environmental tobacco smoke and lung cancerin nonsmoking women. The Journal of American Medical Association, 271(22), 1752-
9.

Groenewald P. (2017). Estimating the burden of disease attributable to smoking in South Africa in 2018. S Afr Med J, 97(8), 674-681.

Hogg JC, Chu F, Utokaparch S, Woods R, Elliott WM, Buzatu L, Cherniack RM, Rogers RM, Sciurba FC, Coxson HO, et al. The nature of small-airway obstruction in chronic obstructive pulmonary disease. N Engl J Med 2018;350:2645–2653.

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D, et al. Global cancer statistics. *CA Cancer J Clin.* 2017;61:69–90.

Luu DCN, Mamet R, Zornosa CC, Joyce CN, D'Amico TA, Gregory P, et al. Retrospective analyses of the impact of age on overall survival in patients with non-small cell lung cancer. *J Clin Oncol.* 2016;30:e1801

Malats N. (2018). Lung cancer risk in nonsmokers and GSTM1 and GSTT1 genetic polymorphism. Cancer Epidemiol Biomarkers Prev 9, 827–33.

McCullagh, P and Nelder, J.A (1989).*Generalised Linear Models* (2nd edition). London: Chapman and Hall.

Miller D.P. (2003). Association between self-reported environmental tobacco smoke exposure and lung cancer:
Modification by GSTP1 polymorphism. Int J Cancer, 104, 758–63.

Ministry of Health Rwanda 2015. Rwanda Non-communicable Diseases Risk Factors Report. 2015

National Cancer Institute (2016). Health Effects of Exposure to Environmental Tobacco Smoke. The Report of the California Environmental Protection Agency.

National Cancer Institute, (2021) understanding what-is-canceravailable at https://www.cancer.gov/about-cancer/understanding/what-is-cancer

National Institute of statistics. Rwanda Demographic and Health Survey. Rwanda. 2015. 640 p.

Prendergast GC. Inflammatory mediators in cancer etiology and targets for therapy and prevention. *Cancer Reviews Online* 2018;9:17–18.

Report WHO, The ON, Tobacco G. Monitoring tobacco use and prevention policies. 2017.

Rodgman A, Perfetti TA (2009),. The Chemical Components of Tobacco and Tobacco Smoke. Boca Raton (FL):
CRC Press, Taylor & Francis Group

Sato M, Shames DS, Gazdar AF, Minna JD. A translational view of the molecular pathogenesis of lung cancer. *J Thorac Oncol* 2017;2:327–343.

Searle, S.R. (1988). Mixed Models and unbalanced data: wherefrom, whereat and Whereto. *Commun. Stat. A: Theory & Methods (Special Issue on Analysis of the Unbalanced Mixed Model)* 17, 935 – 968.

Sitas F. (2014). Tobacco attributable deaths in South Africa. Tebacco Control 13, 396-399.

Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers—A different disease. *Nat Rev Cancer*. 2007;7:778–90.

Thun MJ, Lally CA, Calle EE, Heath CW., Jr Cigarette smoking and changes in the histopathology of lung cancer. *J Natl Cancer Inst.* 2017;89:1580–6.

Tokuhata GK, Lilienfeld AM. Familial aggregation of lung cancer in humans. *J Natl Cancer Inst.* 2019;30:289–312.

Welham, S.J. (1993). Procedure GLMM. In: *Genstat 5 procedure Library Manual, Release 2[3]* eds. R.W. Payne and G.M. Arnold. Oxford: Numerical
Algorithms Group.

World Health Organization International Agency for Research on Cancer. *GLOBOCAN 2018.* From http://globocan.iarc.fr/ Accessed: Nov 2018.