The Spatial Mixed effect Ordinal Logistic Regression Modeling of Cancer disease: West Amhara, Ethiopia

Lijalem Melie Tesfaw (lijalemmelie@gmail.com)
Bahir Dar University College of Science
https://orcid.org/0000-0002-2555-8559

Essey Kebede Muluneh
Bahir Dar University

Research article

Keywords: Incidence of cancer, Patient Status, Spatial (Mixed) OLR

Posted Date: September 7th, 2019

DOI: https://doi.org/10.21203/rs.2.14082/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published on July 29th, 2020. See the published version at https://doi.org/10.1177/1176935120939898.
The Spatial Mixed effect Ordinal Logistic Regression Modeling of Cancer disease: West Amhara, Ethiopia
Lijalem Melie Tesfaw*1; Essey Kebede Muluneh2
Email: *Lijalem Melie Tesfaw=lijalemmelie@gmail.com
Essey Kebede Muluneh=esseykebede@gmail.com
1,2 Department Of Statistics, Bahir Dar University, Po.Box 79 Bahir Dar, Ethiopia.
* Corresponding Author

Abstract

**Background:** There are numerous patients that are suffering with cancer disease in Ethiopia. However, as far as our knowledge concerned there was no any study that have been conducted to determine the factors on cancer disease. The main objective of this study was to assess the distribution risk of cancer disease by taking into account the spatial effect. It was also aimed to determine the effects of some common clinical patients characteristics, complications of cancer disease and prognostic factor on current patient status, and finally predict the patient status.

**Methods:** The data was obtained from 415 cancer patients. Statistical models were used to investigate the spatial difference of incidence of cancer across different districts and different disk factors identified using spatial mixed OLR model.

**Results:** The findings of this study shows that only 1.45% cured of patients who have taken treatment were cured and 46.02% improved while the rest have shown no change and even worse status after they getting treatment. Prognostic factor (stage and grade of cancer tumor), complication of cancer disease such as anemia during diagnosis and treatment of patients given in the hospital had significant effect on the patient status.

**Conclusion:** Patients without anemia complication at diagnosis were more likely to fall in the lower patient category than patients with anemia complication during diagnosis. Most of the patients had advanced stage (IV) and grade of cancer tumor that dismantle the capability of the treatment to be less effective. There was negative spatial effect on the incidence of cancer indicate that districts with higher cancer incidence usually surrounded by districts with lower incidence.

**Keywords:** Incidence of cancer, Patient Status, Spatial (Mixed) OLR
Background

Cancer is a genetic disease caused by changes to genes that control the way our cells function, especially how they grow and divide which leads into unregulated growth and division of cells that forming malignant tumor and invade the nearby parts of our body. The cause of cancer, genetic changes, can be inherited from parents as well as during person’s lifetime as result of errors that occur as cells divide by certain environmental exposures. Cancer tumors are malignant, that can spread into and invade nearby tissue, which is called metastatic cancer, which can start almost anywhere in the human body \([1, 2]\). Recently National Cancer Institute showed that, there are more than 100 types of cancer which usually named for the organs or tissues where the cancers form \([2]\). Among these more than 100 types of cancer, only 20 types of cancers were considered in this study based on the registry of cancer patients in the oncology ward of Felegehiwot Referral Hospital, west Amhara region in Ethiopia.

According to the Global Burden of Disease Cancer result explored in 2015, cancer was the second leading cause of death globally, 8 million deaths, whereas cardiovascular diseases being the first \([3]\). Based on the 2018 world cancer statistics (excluding non-melanoma skin cancer) there were an estimated 18 million cancer case around the world, of these 9.5 million cases were in men and 8.5 million in women. Globally, Lung cancer is the first most commonly diagnosed cancer that affect most cases (12.3% of the total). Breast (2,088,849 cases, 12.2% of the total), Colorectal (10.00%), Prostate (9.8%) and Stomach (5%) are within top five most commonly diagnosed cancer ranked second, third, fourth and fifth respectively \([4]\).

Cancer is a major public health burden in both developed and developing countries. About 72% of all cancer deaths in 2007 occurred in low and middle-income countries \([7]\). Here it is the same pattern of cancer in sub-Saharan African countries, particularly in Ethiopia. In Ethiopia, cancer accounts for about 5.8% of total national mortality. According to national cancer report in 2015 of Ethiopia among the entire adult population breast cancer (30.2%), cancer of the cervix (13.4%) and colorectal cancer (5.7%) are the most prevalent cancers in Ethiopia. Since Ethiopia has diversified geographical area, in this study spatial effect of the cancer disease was also accounted by determining district residence place of the patient. Assigning individuals to their place of residence also poses problems \([30]\), although this is usually the only locational
information that is available. Often the goal of a geographic analysis is to identify a common environmental exposure in a population, but exposures that are occupational or recreational may not necessarily reveal themselves in a residential analysis.

A study in [6] reported that in Africa, number of cancer deaths is rising at an alarming rate that the number of disease from cancer will have increased almost 70% based on age demographics alone in 2030. Most cancers are diagnosed at an advanced stage in Africa mainly because of limited cancer treatment. Over 20% of African countries have no access to cancer treatments at all, while access is limited and sporadic in other countries. There are only four hospitals in Ethiopia that can give treatment for those of patients affected by cancer disease for the time being, which is very limited for a country that has more than 100 million people. Bahir Dar Felegehiwot Referral Hospital is one of the four hospital that can give cancer treatment in Ethiopia which was launched at the end of 2016. The growing population coupled with lifestyle changes will mean an increasing burden of cancer. However, oncology services are wholly inadequate to serve the entire country in the essential time. As a result of this most of the patients are getting treatment at the advanced stage of cancer disease and die within small period or while/as soon as they arrive at the hospital.

The main objectives of this study was to assess the distribution of the risk of cancer disease risk by taking into account the spatial effect over district and explore whether there is spatial autocorrelation of incidence of cancer disease between districts. It was also aimed to determine the effects of some common clinical patient characteristics, complications of cancer disease and prognostic factor on current patient status, and finally predict the patient status.

**Methods**

**Data**

The data in this study was obtained from the Felegehiwot Referral Hospital (FHRH) cancer patients registry in the oncology ward. The oncology ward from the hospital, located in Bahir Dar city, designed to register all cancer patients that comes from different districts of West Amhara region, Ethiopia. The cancer data of patients obtained from the year 2016 September
to 2019 January. Registries attempt to consolidate information by patient Id so that each case appears only once in the registry. For cancer patients series of interventions, including psychosocial support or palliative care, surgery, chemotherapy that is aimed at curing the disease or prolonging life considerably while improving the patients quality of life were made. Data was collected from oncology ward of the hospital retrospectively reviewing all new cancer patients in cancer registry center report. A series of questions using oral and medical tools regarding risk factors, main symptoms, complication, co-morbidities, treatment options (if they used before they came to the hospital) and prevention and early detection measures of cancer were asked to evaluate the cancer’s identity for each patient. After data collection, information was entered into Excel data-sheets and then exported to other statistical software SAS and R. In this study the data have been collected from 415 cancer patients. Of these patients, 285 (68.67%) patients have been taken chemotherapy treatment, 120 (28.92%) patients have been taken both chemotherapy and surgery (combined) treatment while the rest 10 (2.41%) patients have been given palliative care. Palliative care is urgent humanitarian treatment to relieve rather the pain than cure when patients have advanced stages of cancer and little chance of cure.

Variable Description

Before clinical assessment at the oncology ward of Felege Hiwot Referal Hospital (FHRH), patients were asked to complete their life history of patient characteristics (see in Table 1) and unique patient Id assigned. Patient characteristics/risk factors such as: age, gender, residence and blood type; prognostic factors such as: stage and grade; complication such as anemia; treatment and patient status indicated in Table 1 were variables that considered in this study. Variables those mentioned under risk factors are factors associated with causing a cancer disease and determined by looking at things that influence the incidence of new cancer cases, whereas prognostic factors can only be determined by following up people who already have the cancer. Risk factors are patient characteristics associated with the risk of contracting cancer disease which include age, gender, residence and blood type. On the other hand, prognostic factor, complication and treatment were collected from patients during treatment in the hospital and finally the patient status was recorded in their registry card after they received treatment. The prognostic factor and complication of patients registered at the beginning of the diagnosis,
i.e as soon as they came to hospital. Complications of cancer are unanticipated/unforeseen diseases that arise following and as a result of cancer such as anemia in our study. Based on the information of patients in their registry card the following variables were considered in this study (see in Table 1).

**Statistical Modeling**

**Spatial Analysis**

Spatial data is distinguished by observations that are obtained at spatial locations $s_1, s_2, ..., s_i$ where the $s_i$ are coordinates in the plane $\mathbb{R}^2$ and rarely in the space $\mathbb{R}^3$. Spatial data have different features such as: point, line, area and volume. Point: a precise location, $s$, in space indicated by a dot on a map; Line: a sequential collection of connected points like road and rivers; Area: a region enclosed by lines like counties, states and districts, one features of spatial data that have been considered in this study. Finally, volume is a spatial data feature of object with three dimension which is common for geologic formation.

In spatial data analysis, locations close together in space often have similar values of outcome variables while locations far apart are often different. Everything is related to everything else, but near things are more related than far thing. This law succinctly defines the statistical notion of (positive) spatial autocorrelation, in which pairs of observations taken nearby are more alike than those taken farther apart. This spatial correlation must be taking into account for spatial analyses. Spatial analysis is an analysis which includes the influence of space into the analysis. All statistical methods for spatial data have to take the spatial arrangement, and the resulting correlations, of the observations into consideration in order to provide accurate and meaningful conclusions based on the analysis.

**Weighted matrix to test Spatial autocorrelation**

Weighted matrix, sometimes called Contiguity matrix, describes the relationship between districts $i$ and $j$ in the specified area. The $(i, j)^{th}$ element of a spatial proximity matrix $W$, denoted
$w_{ij}$, quantifies the spatial dependence between regions $i$ and $j$, and collectively, the $w_{ij}$ define a neighborhood structure over the entire area. The spatial correlation parameter, and $W=(w_{ij})$ is a neighborhood matrix for the areal units [9], which can be defined as

$$w_{ij} = \begin{cases} 1, & \text{if districts } i \text{ and } j \text{ share a common boundary }, i \neq j \\ 0, & \text{Otherwise} \end{cases}$$  \hspace{1cm} (1)$$

In this case the symmetric properties of $W$ are established because of that $w_{ji} = w_{ij}$ and its diagonal elements equal to zero, being the similarity of the $i^{th}$ region with itself $w_{ii} = 0$.

Spatial autocorrelation, covariation or correlation between neighboring observations of variables, gives us information on similarity between observation and similarity among locations. Based on the collection of weighted matrix, spatial proximity matrix, there are different measures of similarity that define different index classes. The two standard statistics that are used to measure strength of spatial association among areal units are Moran’s I and Geary’s C [9]. Moran’s I is widely used [11], and variations of it relate to likelihood ratio tests and best invariant tests for particular models of correlation for normally distributed random variables. Having the spatial proximity matrix constructed, the Moran’s I statistics as a measure of global indexes of spatial autocorrelation with spatial proximity and similarity between areas $i$ and $j$ can be written as follows:

$$I = \frac{1}{S^2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (Y_{ij} - \bar{Y})(Y_{j} - \bar{Y}) = \frac{1}{S^2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} Y_{ij}$$  \hspace{1cm} (2)$$

Where $Y_i$ refers to the total number of incidence of cancer diagnosed in the $i^{th}$ district, $\bar{Y} = \sum_{i=1}^{N} Y_i / N$ is the overall mean and $S^2$ is the sample variance observed in the $Y_i$’s that can be computed as $S^2 = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$.

In test of spatial significance/autocorrelation the null hypothesis states that the near-by districts do not affect one another which implies that there is independence and spatial randomness in the data. In contrast, the alternative hypothesis states that there is spatial association or dependence among the districts. The research hypothesis for this study states that the near-by districts in West Amhara region have an association or dependence on the disease risk of cancer. Spatial autocorrelation in the near-by areas is considered to be present when the test statistic
such as Moran’s I takes on a larger value, compared to what would be expected under the null hypothesis of no spatial association [9][11].

The significance of spatial autocorrelation can then be detected by comparing the statistic \(I\) with the expected value of \(I\), 

\[ E(I) = -\frac{1}{N-1} \]

In case of randomization assumption, data values are reassigned among the \(N\) fixed locations, providing a randomization distribution against which we can judge our observed value. When \(I\) lies in the tails of this distribution, we reject the assumption of independence among the observation and, hence, significant spatial autocorrelation in the data. Therefore, if we rely on the normality assumption beyond the expected of \(I\), \(E(I)\), it is also required to compute the variability of \(I\), \(\text{Var}(I)\), so that we can calculate 

\[ Z = \frac{I - E(I)}{\sqrt{\text{Var}(I)}} \]

The variance of \(I\), \(\text{Var}(I)\), can be computed as given in Equation 3 given by:

\[ \text{Var}(I) = \frac{N^2 S_1 - NS_2 + 3S_0^2}{(N - 1)(N + 1)S_0^2} - \left(\frac{1}{N - 1}\right)^2 \]  

(3)

where, \(S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}\), \(S_1 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (w_{ij} + w_{ji})^2\), and \(S_2 = \sum_{i=1}^{N} (w_{i+} + w_{+j})^2\), with \(w_{i+} = \sum_{j=1}^{N} w_{ij}\) and \(w_{+j} = \sum_{i=1}^{N} w_{ij}\). Given this, the decision of the hypothesis of the spatial autocorrelation can be made by comparing the z-score statistic with corresponding standard normal distribution value. The Moran’s I statistic which is greater than the expected value \(E(I)\) points to a positive spatial autocorrelation (clustered pattern), while a value of Moran’s I that is below the expected value indicates a negative spatial autocorrelation (regular pattern).

However, using Moran’s I statistic for count data, i.e. incidence of cancer per districts in our case, assess the spatial similarity of deviations of each regional count \(Y_i\) with the overall mean district count \(\bar{Y}\) in Equation 2 is very doubtful and under questionable to really assess the clustering/spatial autocorrelation. This is because spatial heterogeneity of regional at-risk population sizes inherent in districts public health data, observed spatial similarity in regional deviations from the mean regional count may simply be due to variations in the district at-risk population size. For instance, extremely small numbers in the lower range of cases are observed in very small districts, such as Injibara and Finote Selam, and high values partly attributed to highly populated districts, such as Bahir Dar. So, this may suggest that the observed spatial similarity in district deviations from the average district count may simply be due to variations in the regional at-risk population size. To handle such problem, i.e. to standardized the population at risk, the incidence counts were replaced by incidence proportions, to remove the impact of
population heterogeneity in some amounts [1].

Spatial Variable, $W_{ij}Y_{ij}$, is a variable with a product of weighted matrix $W_{ij}$ and $Y_{ij}$, a cross product of values in district i and j that deviated from average value $\bar{Y}$, Equation 2. $Y_{ij}$ is needed to measure the proximity which means the distance between the observed i value and neighboring j values. Though are several types of distance based methods [31], the most common distance method is an Euclidian distance, $Y_{ij} = (Y_i - \bar{Y})(Y_j - \bar{Y})$ was used for this study.

**Ordinal Logistic Regression Model**

The logistic regression model is a model used to study the association between a categorical dependent variable and any set of independent variables [28] with any data type and ordinal logistic regression when a dependent variable has only two values, more than two values and more than two values with having natural order or rank respectively. As the dependent variable in this study, patient status, belongs to a variable that has five ordinal values, ordinal logistic regression model was used to analyze the data.

Ordinal logistic regression analysis deals with the association of a dependent variable with independent variables when the dependent variable has more than two categorizes having natural order or rank. The dependent variable $Y$ is assumed to have an ordinal scale with $J$ categories and $X = (x_1, x_2, ..., x_p)$ is the vector of explanatory variables. Then the chances of the variable response of the j-th category of explanatory variable x in particular can be expressed by $P$, $P[Y=j|x] = \pi_j(x)$.

When response categories are ordered, the logits can utilize the ordering that results greater power and simple interpretation. The cumulative probability for $Y$ is the probability that $Y$ falls at or below a particular outcome category $j$ and is given by:

$$P(Y \leq j) = \pi_1(x) + ... + \pi_j(x)$$

, $j = 1, 2, ..., J$. Where J is number of categories for the response variable $Y$. 

The cumulative logit model [28][25] is given in Equation 5 as follows:

$$
\text{logit}(P[Y \leq j|x]) = \log\left(\frac{P[Y \leq j|x]}{1 - P[Y \leq j|x]}\right) = \log\left(\frac{\pi_1(x) + \pi_2(x) + \ldots + \pi_j(x)}{\pi_{j+1}(x) + \pi_{j+2}(x) + \ldots + \pi_J(x)}\right) = \alpha_j + X'\beta
$$

, j=1, 2, ..., J-1 and the probability $P[Y \leq j|x]$ can be estimated as:

$$
P[Y \leq j|x] = \frac{\exp(\alpha_j + X'\beta)}{1 + \exp(\alpha_j + X'\beta)}
$$

The cumulative probabilities do not use the final one, P(Y ≤ J), since it necessarily equals 1. The parameter $\beta$ is a vector of regression coefficients describing the effect of the corresponding independent variable $X$ on the log odds of response in category $j$ or below. When this model fits well, it requires a single parameter rather than J-1 parameters to describe the effect of $X$. Because the model assumes that the effect of $X$ is identical (proportional odds) for all J-1 cumulative logits.

**Mixed ordinal logistic regression model**

The term mixed model refers to the use of both fixed and random effects in the same analysis. Fixed effects have levels that are of primary interest and would be used again if the experiments were repeated [33]. Random effects have levels that are not of primary interest, but rather are thought of as a random selection from a much larger set of levels. In this study, patients and their corresponding cancer type were considered as random effects because we realized that the cancer types included in this study might not be the only cancer types if the cancer type registry inspect in the study area beyond to the hospital, FHRH. Hence, the term mixed logistic regression model revealed ordinal logistic regression model that consists both random and fixed effects/factor. Mixed ordinal logistic regression model is one of a generalized linear mixed models (Glimmix), models with a particular cumulative logit link. The general linear mixed model [32] is given by:

$$
g(\mu_i) = X_i'\beta + Z_i'\mu
$$
\( \mu_i = \text{E}[y_i / \mu] \), where \( g(.) \) is the link function since it links together the conditional mean of \( y_i \) and the linear form of predictors. \( X_i \)'s are independent variables and \( \beta \) the parameter that manifest corresponding effect of independent variables. \( Z_i \)'s are random effects and \( \mu \) the corresponding random effect parameter vector. \( \mu_i \) represents the conditional mean rather than the marginal or unconditional mean; otherwise, all is the same.

In this study the link function \( g(.) \) is the logit link and the model of the response variable \( Y_{ik} \) of patient status can be build as:

\[
\logit(P[Y_{ik} \leq j|x, \mu_k]) = \log \left( \frac{P[Y_{ik} \leq j|x, \mu_k]}{1 - P[Y \leq j|x, \mu_k]} \right) = \alpha_j + X'\beta + \mu_k
\]

, \( j=1, 2, ..., J-1 \). where \( \mu_k \) is the random effect of the \( k^{th} \) cancer type or patient ID and \( X'\beta \) is denote fixed factors mentioned in the variable description in Table 1.

**Random effect test**

Random effects model is commonly used to detect whether the variable intended to have random effect or not, beyond to the fixed effect that is encompassed in the model [32] [33]. Random factors is not restricted to linear mixed models, researchers want to incorporate random factors into nonlinear models to build a model that accommodates correlated data, or to consider the levels of a factor as selected from a population of levels in order to make inference to that population [32]. Common questions in mixed modeling are whether variance components are zero, whether random effects are independent, and whether rows (columns) can be added or removed from covariance matrix. The effect of cancer type and patient ID were detected by including in the model as a random effect. The patient status within one cancer type are likely to be correlated each other. The goal here is to make inference for the population in the study area of cancer types. This could be accommodated by incorporating cancer type and patient ID into the model.

The random test can be conducted in two ways: i) adding the random effect in the model that consists the fixed effects and perceive if there is pragmatic change in the estimated parameter or ii) using mixture chi-square statistic [28] [32].
Spatial Mixed Ordinal Logistic Regression

Spatial Mixed Ordinal Logistic Regression (SMOLR) is an analysis which incorporates spatial effects into mixed ordinal logistic regression model. Scholars in [24][25] used different methods to account spatial effect and estimate its effect in their model. The spatial logistic regression model in [25] accounted for the spatial effect by including weighting of the location of the $i,j^{th}$ is through $w_{ij}=1/h_{ij}$ where $h_{ij}$ is the euclid distance between the district $i, j$ and estimated existence of event. On the other hand, spatial mixed ordinal logistic regression model is established to handle the spatial relationship proposed by [24] and the specific model used in this study is given by:

$$logit(P[Y_{ik} \leq j | x, \mu_k]) = \log \left( \frac{P[Y \leq j | x, \mu_k]}{1 - P[Y \leq j | x, \mu_k]} \right) = \alpha_j + X\beta + \rho W y + \mu_k$$  \hspace{1cm} (9)

where $j$ is the $j^{th}$ category of the dependent variable and $W$ is a weighting matrix that represents the spatial proximity of the region, $w_{ij}=0$ if regions $i$ and $j$ are not adjacent districts directly while $w_{ij}=1$ if districts $i$ and $j$ areas are immediately adjacent. Spatial weighting matrix ($W$) which has been obtained is multiplied by the vector $y$ and the results will be considered as new variable the so called spatial variable and will be used in ordinal logistic regression analysis [24]. The parameter vectors $\beta$ and $\rho$ refer to the effect of the independent and spatial variable respectively.

Goodness of fit of the model

Before making any prediction we need to select the best model out of the proposed spatial ordinal logistic regression (SOLR) and SMORL models, the likelihood and pousdo likelihood methods were used to estimate the parameters in the SOLR and SMOLR models respectively. To choose the best model Bayesian information criterion (BIC) and Akaike information criterion (AIC) were used [28] [25]. The validity of the selected model will then be checked using the Correct Classification Rate (CCR) [25]. CCR is the percentage of correct observations (suitability with
the expected value) which is given by:

\[ CCR = \frac{\text{number of correct prediction}}{\text{number of observation}} \times 100\% \]  

(10)

When the number of correct prediction is high, the Correct Classification Rate CCR become high. Hence, the higher percentage of CCR shows higher accuracy.

**Results**

**Exploratory Data Analysis**

Prior to fitting the model, the nature of the data that could be used as a guide for the modeling framework was examined. The sample size within each cancer type was inadequate to make separate analysis of factors within different cancer type. Hence, in this study data exploration was done in general for cancer irrespective of cancer type. To see the effects of different factors on patient status, it is advisable to first explore the incidence of cancer for detail insight about the data. Then after a bivariate association between all categorical covariates and patient status was explored so that whether significant association of covariates with patients status can be detected.

Most of the patients (104 cases, 25.06% of the total) had advanced stages (IV) of cancer during diagnosis. This late detection and treatment of cancer leads to greatly aggravate the burden of cancers and bad outcomes of patients status. All patients took appropriate treatment based on Physicians prescription. Of those patients who have taken treatment, only 1.45% were cured and 46.02% have improved while the rest have no change because of treatment and even worse after getting the treatment. This shows that it is less likely that the patient can be cured, but that the diseases effect can be improved if effective measures are put in place to control risk factors, detect cases early and offer good care to those with the disease. The response variable, patient status, is constructed by categorizing the severity of cancer disease into five ways by cured, improved, same, deteriorate and death and the percentage distribution presented in Table 1.

This type of categorization is used by the hospital to manifest the general status of patients.
after treatment.

Table 1: Variables description considered in the study

| Independent variables | frequency(%) |
|-----------------------|--------------|
| *Patient characteristics/risk factors: | | |
| **Age** | | |
| **Gender** | | |
| female(0) | 257 (61.93) |
| male(1) | 158 (38.07) |
| **Residence** (in districts, Table 3) | | |
| **Blood type** | | |
| A+ | 106 (25.54) |
| A- | 18 (4.34) |
| AB+ | 17 (4.10) |
| AB- | 4(0.96) |
| B+ | 101 (24.34) |
| B- | 16 (3.86) |
| O+ | 121 (29.16) |
| O- | 32 (7.71) |
| *Prognostic factor: | | |
| **Stage grade** | | |
| I | 36 (8.67) |
| II | 68 (16.39) |
| III | 84 (20.24) |
| IV | 104 (25.06) |
| low | 100 (24.10) |
| high | 23 (5.54) |
| *Complication: | | |
| **Anemia at diagnosis** | | |
| 1=yes | 227 (54.70) |
| 0=no | 188 (45.30) |
| *Treatment: | | |
| **Treatment** | | |
| Chemotherapy | 285 (68.67) |
| Chemotherapy and surgery | 120 (28.92) |
| Palliative care | 10 (2.41) |
| Dependent variable | | |
| *Patient Status: | | |
| **Patient status** | | |
| 0=cured | 6 (1.45) |
| 1=improved | 191 (46.02) |
| 2=same | 108 (26.02) |
| 3=deteriorate | 85 (20.45) |
| 4=death | 25 (6.02) |

key: Stage_grade= Stage (I, II, III, IV) and Grade (high, low)

Number of patients distribution for male and female within different cancer types are different as it revealed in Figure 1. It displayed in Table 3 that a large proportion of female patients
suffer from breast, gastric, HCC, rectal and ovarian cancer. Especially in breast cancer there is huge predominant difference of female over male. On the other hand, colonic, HL, NHL, pancreas, RBCT and sarcoma cancer males patients were more predominant on males than female patients.

There were 20 cancer types available and considered in the study based on the oncology ward cancer registry at the Felege Hiwot Referral Hospital, Table 2 below. The Table for cancer types revealed all types of cancer in the study and their corresponding number of patients (in percent %) that belongs into each cancer type among a total of 415 patients. Breast Cancer was responsible for the highest diagnosed cancer type (19.28% of the total) followed by NHL Cancer (16.87%).

| Cancer type  | Lung | NHL | Breast | Cervical | Colonic | Endometrial | Esophageas |
|--------------|------|-----|--------|----------|---------|-------------|------------|
| No. of patients | 29  | 71  | 80     | 52       | 10      | 2           | 5          |
| Percent(%)    | 6.99 | 16.87 | 19.28 | 12.53    | 2.41    | 0.48        | 1.20       |

| Cancer type  | HCC | HL | Ovarian | Pancreas | RBCT | Rectal | Sarcoma |
|--------------|-----|----|---------|----------|------|--------|---------|
| No. of patients | 12  | 7  | 36      | 12       | 6    | 13     | 50      |
| Percent(%)    | 2.89 | 1.69 | 8.67    | 2.89     | 1.45 | 3.13   | 12.05   |

| Cancer type  | Testicular | head& neck | Nasophyrege | Skin | CLL | Gastric |
|--------------|------------|------------|-------------|------|----|---------|
| No. of patients | 4   | 7          | 2           | 2    | 7  | 9       |
| Percent(%)    | 0.96 | 1.69       | 0.48        | 0.48 | 1.69 | 2.17    |

Forty one districts from six zones were considered in the study, Table 3. There are thirteen, eight, seven and ten districts from west Gojjam, East Gojjam, Awi and South Gondar zones respectively, while both South wollo and Benshangual Gumz are the only zones that consists only one district. Those districts are not districts selected by ourselves rather they are the only districts that patients were available in oncology cancer registry ward in the hospital.
Address-matching of cancer registry records was obtained from zone administration of the districts that have been done using a commercial GIS product to assign approximate latitude-longitude coordinates to a patient’s reported district place of residence during diagnoses of a cancer tumor. There were differential variations in patterns across the forty one districts. The highest incidence of cancer was recorded in Bahir Dar. However, population at risk is not in the standardized form, in a sense that the highest incidence of cancer in Bahir Dar might be due to high number of inhabitants in the district/city. This suspected problem was handled by
standardizing the population at risk per 100,000 per district. Finote Selam district was then found to have the highest incidence of cancer relative to other districts in the study, Table 4. Mean age of patients at diagnosis was 43.5 years with standard deviation value 15.54 and median age was 45. The minimum age of patients included in the study was 2 years while the maximum was 82. Most of the patients (72%) were between 30 and 60 years old. On the other hand, the rest 16% of patients were less than 30 years old while 12% of patients were over 60 years old.

Even if the population at risk within each age group were not standardized, the incidence count of cancer disease for females and males looks different as it presented in Figure 2 below in which the incidence count for females were more peaked than males between age group 21 to 70 years and vice versa for age group below 21 and above 70 years old. Figure 2 shows the incidence count of cancer within different age groups irrespective of sex of patients. The incidence count of cancer disease increase with increasing rate till age 50, using the average cubic estimated reference line. After age 50 the incidence count of cancer disease become decline with increasing rate.

For the sake of simplicity to explore the incidence count of cancer with age, the variable was constructed by categorizing into different age groups in which each group contains five years. This type of categorization is consistent with previous report [2]. In the figure, the incidence count of cancer in each age group was obtained by summing up number of patients within the specified age group.

The description of incidence of cancer disease that described in Table 3 did not account for the population at risk for each district that leads into biased conclusion. The highest incidence of cancer disease was recorded in Finote Selam district with incidence rate 19.5929 followed by Chagni district with incidence rate 14.6985 while the lowest incidence of cancer was recorded from Enemay districts with incidence rate 1.1117 and followed by Smada district with incidence rate 1.2120. Hence, in order to determine the incidence of cancer for each district, the population at risk should first be considered to have plausible description of the incidence rate. The latitude (lat.) and longitude (lon.) which shows where each district is located in the western part of Amhara region was illustrated graphically in Figure 3 which shows the incidence of cancer within each district.
Table 4: Incidence of cancer disease for per 100,000 standardized population at risk for each district

| District          | lon. | lat.   | Inc. of cancer | District          | lon. | lat.   | Inc. of cancer |
|-------------------|------|--------|----------------|-------------------|------|--------|----------------|
| North Achefer     | 11.60| 37.03  | 10.6319        | Ankasha Guagusa   | 11.00| 36.67  | 6.9079         |
| South Achefer     | 11.83| 37.17  | 8.2558         | Banja             | 11.17| 36.25  | 7.6073         |
| Burie             | 10.70| 37.06  | 5.7549         | Chagni            | 10.95| 36.50  | 14.6985        |
| Bahir Dar         | 11.60| 37.38  | 6.7459         | Dangla            | 11.42| 36.67  | 7.5627         |
| Bahir Dar Zuria   | 11.25| 37.17  | 12.5115        | Fagita Lekoma     | 11.33| 36.75  | 3.6430         |
| Dembecha          | 10.67| 37.17  | 7.0974         | Guangua           | 11.00| 35.25  | 1.6448         |
| Finote Selam      | 10.71| 37.26  | 19.5929        | Jawi              | 11.75| 36.42  | 6.9750         |
| Gonji Kolela      | 11.45| 37.67  | 8.6465         | Alefa             | 12.25| 36.33  | 2.7900         |
| Mecha             | 11.50| 37.00  | 7.8723         | Libo Kemkem       | 12.33| 37.68  | 6.0058         |
| Merawi            | 11.25| 36.50  | 13.3984        | Debre Tabor       | 11.85| 38.22  | 7.9379         |
| Sekela            | 11.17| 37.00  | 3.9887         | Dera              | 11.75| 37.50  | 6.7287         |
| Woberma           | 10.37| 35.75  | 4.5338         | Estie             | 11.67| 38.17  | 2.7861         |
| Jabi Tehnan       | 10.80| 37.17  | 7.7315         | Ebinat            | 12.17| 38.08  | 3.8754         |
| Yilmanadensa      | 11.50| 37.33  | 6.8592         | Farta             | 12.00| 38.00  | 2.3829         |
| Bibugn            | 11.00| 37.58  | 6.7994         | Fogera            | 11.97| 37.68  | 3.6484         |
| Debre Markos      | 10.23| 37.32  | 7.0087         | Gaynt             | 12.00| 38.33  | 1.4902         |
| Enebse Sarmidr    | 11.08| 38.25  | 3.4099         | Smada             | 11.50| 38.25  | 1.2120         |
| Enemay            | 10.67| 38.00  | 1.1117         | Dessie            | 11.13| 38.53  | 3.7800         |
| Goncha Siso enese | 11.17| 38   | 3.1030         | Pawi              | 11.33| 36.33  | 6.2724         |
| Hulet ejun enese  | 11.25| 37.75  | 3.6443         | Mota              | 11.08| 37.87  | 12.9493        |
| Machakel          | 10.67| 37.33  | 3.1184         |                   |      |        |                |

The categorical covariates have different number of categories based on behaviors of the covariate. The bivariate analysis using chi-square statistic test is presented in Table 5. This bivariate table shows the relationship between a categorical independent variable and the dependent variable (patient status of categories cured, improved, same, deteriorate and death). Chi-square statistic was used to compare the actual frequencies in a bivariate table with the frequencies that would have been expected if there was no relationship between the variables. We can see...
from the table that the only variables not significantly associated with patient status were sex and blood group, meaning that the patterns of responses across both sex and different blood groups were essentially the same for all categories of the independent variable.

Table 5 shows that there is no any cancer patients with anemia complication who was cured. This shows how a complication of anemia on cancer would aggravate the severity of cancer. In Parallel, of those patients who died of cancer higher proportion of patients (68.0%) were with anemia and the remaining (32%) without anemia. From this we can highly suspect that the patient status have a significant association with anemia. The p-value on the last column found to be 0.001 that indicates there a significant association of anemia with patients status.

The patterns of proportion of cancer patients with and without anemia complication seems to vary by patients status. Patients without anemia complication were more likely than patients with anemia complication to be entered for the lowest scale patient status (100% vs 0.00% for cured and 61.8% vs 38.2% for improved patient status) and less likely than patients with anemia complication to be entered for the highest scale patient status (44.7% vs 55.3% for deteriorate and 32.0% vs 68.0% for death patient status). The reverse is true for patient with vs without anemia complication.

There was significant association between (p-value=0.001) treatment given for patients and patient status. Among patients who have same patient status 75.9% of the patients took chemotherapy while 20.4% of patients with same status have taken both chemotherapy and surgery (combined). However, stage grade of patients have a significant association with patient status. Most of the patients who died of cancer were in Stage IV and in Stage III (60% and 16% respectively). This shows that the odds of dying of cancer increase with the stage of patients, meaning the patients in the advanced stages are more likely to die.
## The Spatial Mixed effect Ordinal Logistic Regression Modeling of Cancer disease: West Amhara, Ethiopia

Table 5: Frequency distribution of categorical covariates by patient status classification and Bivariate analysis

| Covariates | Patient Status | n(%) | P-value |
|------------|----------------|------|---------|
|            | cured          | improved | same | deteriorate | death |
|            | n(%)           | n(%) | n(%) | n(%) | n(%) | |
| Anemia     | no             | 6(100) | 118(61.8) | 57(52.8) | 38(44.7) | 8(32.0) | 0.001 |
|            | yes            | 0(0.0) | 73(38.2)   | 51(47.2) | 47(55.3) | 17(68.0) | |
| Blood Group| A-             | 0(0.0) | 8(4.2)     | 5(4.6)   | 5(5.9)   | 0(0.0)   | |
|            | A+             | 1(14.29)| 47(24.6)   | 24(22.2) | 25(29.4) | 9(36.0)  | 0.89 |
|            | AB-            | 0(0.0) | 2(1.0)     | 2(1.9)   | 0(0.0)   | 0(0.0)   | |
|            | AB+            | 0(0.0) | 8(4.2)     | 5(4.6)   | 3(3.5)   | 1(4.0)   | |
|            | B-             | 0(0.0) | 7(3.7)     | 3(2.8)   | 6(7.1)   | 0(0.0)   | 0.99 |
|            | B+             | 1(14.29)| 49(25.7)   | 26(24.1) | 19(22.4) | 6(24.0)  | |
|            | O-             | 2(28.57)| 13(6.8)    | 7(6.5)   | 9(10.6)  | 1(4.0)   | |
|            | O+             | 3(42.86)| 57(29.8)   | 36(33.3) | 18(21.2) | 8(32.0)  | |
| Treatment  | Chemotherapy   | 3(50.0) | 124(64.9)  | 82(75.9) | 58(68.2) | 18(72.0) | |
|            | Combined       | 2(33.3) | 66(34.6)   | 22(20.4) | 26(30.6) | 4(16.0)  | 0.001 |
|            | pallative      | 1(16.7) | 1(0.5)     | 4(3.7)   | 1(1.2)   | 3(12.0)  | |
| Sex        | female         | 3(50.0) | 112(58.6)  | 65(60.2) | 59(69.4) | 18(72.0) | 0.346 |
|            | male           | 3(50.0) | 79(41.4)   | 43(39.8) | 26(30.6) | 7(28.0)  | |
| Stage Grade| high           | 5(83.3) | 56(29.3)   | 29(26.9) | 7(8.2)   | 3(12.0)  | |
|            | low            | 0(0.0)  | 10(5.2)    | 9(8.3)   | 3(3.5)   | 1(4.0)   | |
|            | I              | 1(16.7) | 18(9.4)    | 8(7.4)   | 9(10.6)  | 0(0.0)   | 0.000 |
|            | II             | 0(0.0)  | 46(24.1)   | 16(14.8) | 4(4.7)   | 2(8.0)   | |
|            | III            | 0(0.0)  | 50(26.2)   | 21(19.4) | 9(10.6)  | 4(16.0)  | |
|            | IV             | 0(0.0)  | 11(5.8)    | 25(23.1) | 53(62.4) | 15(60.0) | |

key: Combined= Both chemotherapy and Surgery; Stage grade= Stage (I, II, III, IV) and Grade (high, low).

### Spatial autocorrelation, Moran’s I test

The weighted matrix W was defined to evaluate the existence of spatial autocorrelation. The matrix allows the measurement of the non-random association between the value of incidence...
The Spatial Mixed effect Ordinal Logistic Regression Modeling of Cancer disease: West Amhara, Ethiopia

of cancer observed in a given district unit and the value of incidence of cancer observed in neighboring district units which used to compute Moran’s I index. In Table 6, the Moran’s I index is univariate analysis that measures the spatial autocorrelation, in a sense that the index allows to detect the incidence of cancer in a given district may be similar to those neighboring districts. The estimated Moran’s I statistic index is \(-4.91 \times 10^{-2}\) with p-value values 0.0022. The p-values are found to be less than 0.05 suggesting significant evidence of unexplained spatial autocorrelation in the incidence of cancer and the negative Moran’s I values indicate negative spatial association. On the other hand, to identify whether there is spatial correlation, beyond to having look the p-value, simply we can compare observed Moran’s I statistic index with expected Moran’s statistic index \(E(I) = -\frac{1}{n-1}\). The observed value of Moran’s I = \(-4.91 \times 10^{-2}\) is less than its expected value \(E(I) = -2.4 \times 10^{-3}\), showing a negative spatial autocorrelation or clustered pattern, which means that observations from nearby district tend to be less alike than observations from districts farther apart.

Table 6: Indicator of spatial autocorrelation

| Indicator | Statistic | p-value |
|-----------|-----------|---------|
| Moran’s I | \(-4.91 \times 10^{-2}\) | 0.0022 |

To make the incidence of cancer comparable among different districts, the population at risk was standirdized into 100,000 population. As a result the values of incidence of cancer indicates the incidence of cancer out of a total of 100,000 population at risk in the given districts.

The Moran’s I index used to test the spatial autocorrelation can be described graphically (Figure 3) and can be sure whether the decision made in the Moran’s I statistic was plausible and persistent. It also verify whether there is evidence for clustering of incidence of cancer, or whether there are districts that have unusual clusters of cancer so that simply we can perceive causal explanations by seeing if there is an spatial patterning of incidence of cancer.

In Figure 3, each circle point a district and the color from warmer blue color to warmer red color shows districts with lower to higher incidence of cancer disease respectively. Extremely high value of incidence of cancer (19.5929) was found in Finote Selam district and the extremely low incidence of cancer (1.1117) was found in Enemay district as described (Table 4). Enemay district found to be the lower incidence of cancer which might be because of the society who
lives in this district are far from the hospital.

Model fitting and estimating parameters

Prior to directly fitting the model some preliminary analysis was done to select the best fit model with appropriate covariates. The spatial variable \( (S_v) \) was created using a multiplication of spatial weighting matrix \( W \) with incidence of cancer \( (Y) \) in each district simply can be denoted by \( S_v \) instead of \( W_y \). The spatial weighting matrix \( W \) looking at the closeness between districts which contains 1 if adjacent and 0 if not directly adjacent, see Equation 1. Hence, for model building spatial variable \( (S_v) \) was used instead of districts.

Variables were selected using stepwise method of selection. In fact, it was also done manually using backward and forward selection method which results same selected variables. During stepwise variable selection, the selection was done by 15% entry and 20% stay significance level, as it also mentioned in the study \[28\]. Across selected variables all possible pairwise interaction effects were examined in the model. However, there wasn’t any interaction effect between all selected covariates with 15% entry and 20% stay significance level. Since numerous studies in different area and era \[13\]-\[18\] reported that the cancer disease is highly interrelated with age, we realize that keeping covariate age in our model (Table 7) is advisable. To perceive the effect of independent variables at the beginning intercept only OLR model without any independent variables was fitted. Then a model that includes all independent variables including the spatial variable. Spatial OLR model was fitted and the model comparison was done using AIC and BIC. The estimated fit statistic of AIC (BIC) is 1056.067 (1072.180) and 924.848 (981.244) for intercept only OLR and spatial OLR model respectively. The dummy variable and the reference (*) variable indicated in Table 7 and the estimated spatial OLR model can be written by:

\[
\logit(P[Y_i \leq j]) = \alpha_j + \beta_1 Age_i + \beta_20 Anemia_i + \beta_30 chemotherapy_i + \beta_31 combined_i + \beta_40 Stage\_grade_{1i} + \beta_41 Stage\_grade_{11i} + \beta_42 Stage\_grade_{111i} + \beta_43 Stage\_grade_{1IVi} + \beta_44 Stage\_grade_{high}\}
\]

Where \( \alpha_j \) is intercept and \( j=0,1,2 \) and 3 for cured, improved, same, deteriorate status of patient respectively and \( i=1,2,\ldots,415 \). While \( \beta_1, \beta_2, \ldots, \beta_m \) are parameters for corresponding
independent variables to be estimated and describe the effect of the corresponding variable on
the log odds of patient status at \( j \) or below category. In the parameter \( \beta \)'s there is no subscript
\( j \) because of the model assumes proportional odds, the effect of independent variables on the
patient status is identical for all \( J-1 \) cumulative logit model.

The requirement of random effects of cancer type into the model was tested using a mixture
of chi-square distribution. These tests were done by adding cancer type as random effect in
the Spatial OLR model which become Spatial mixed OLR model. The need for random effect
was significant (\( \chi^2_{0.1}=6.28; \) p-value=0.0254), which indicates that the model without cancer type
random effects does not fit the data well. The estimated variance of random effects of cancer type
was 0.2997 by assuming constant variance covariance working correlation structure. However,
the estimate variance for patient ID was 0, indicates that the level between patient variability
is not sufficient to warrant incorporating as random effects in the model. Therefore, since there
is no variation, using Patient ID as random effect is unnecessary.

To estimate parameters of Spatial mixed OLR model estimation methods such as commutative
logistic mixed model (clmm) and generalized linear mixed model (glimmix) using R and SAS
statistical software respectively were used. Nevertheless, at the end the estimated parameters
for the models described in Table 7 we used glimmix SAS procedure.

In case of model comparison based on AIC and BIC, the model with the smallest AIC and BIC
is preferred. The fit statistics AIC (BIC) for intercept only OLR model, spatial OLR model
and spatial mixed OLR were 1056.067 (1072.180), 924.848 (981.244) and 923.040 (937.970)
respectively that revealed in Table 7. Hence, since spatial mixed OLR model has the smallest
AIC (BIC), it is predominantly best model relative to intercept only OLR and spatial OLR
model. In terms of AIC the spatial OLR model and spatial mixed model are close to each other
but in terms of BIC. On the other hand, even the fit statistics in terms of BIC is different
for the two models there is no any pragmatic significance difference of estimated parameters of
fixed covariates but it does matter in case of prediction since the estimated constant for the two
models are quite different. Hence, in the mean time, the inferential discussion was carried out
using spatial mixed OLR model by checking the goodness of fit of the model.
Table 7: Parameter estimates using Intercept only OLR model, Spatial OLR model and Spatial mixed OLR model.

| Parameter | Intercept Only OLR model | Spatial OLR model | Spatial mixed OLR model |
|-----------|--------------------------|-------------------|------------------------|
|           | Estimate (se.err) | p-value | Estimate (se.err) | p-value | Overall p-value | Estimate (se.err) | p-value | Overall p-value |
| Intercept |              |         |              |         |                |              |         |                |
| cure      | -4.222 (0.4112) | <.0001  | -4.3954 (0.5702) | <.0001  | -6.2469 (0.9357) | <.0001  |
| improved  | -0.1013 (0.0983) | 0.3028  | 0.1292 (0.4159)  | 0.7561  | -1.6466 (0.8322) | 0.0625  |
| same      | 1.0198 (0.1112)  | <.0001  | 1.6861 (0.4242)  | <.0001  | -0.03957 (0.8276) | 0.9624  |
| deteriorate | 2.7473 (0.2063) | <.0001  | 3.8532 (0.4754)  | <.0001  | 2.1793 (0.8518)  | 0.0192  |
| death*    | 0.0000         |         | 0.0000         |         | 0.0000         |         |
| Age       |              |         | -0.0028 (0.0064) | 0.6612  | -0.0030 (0.0067) | 0.6570  | 0.6570  |
| Anemia    |              |         |              |         |                |              |         |                |
| no        |              |         | 0.3270 (0.0989) | 0.0009  | 0.0000         |         |
| yes*      |              |         | 0.0000         |         | 0.0000         |         |
| Treatment |              |         | 0.3329 (0.2194) | 0.1292  | 1.4549 (0.6084) | 0.0173  |
| Chemotherapy |              |         | 0.8145 (0.2409) | 0.0007  | 0.0031         | 0.0311  |
| Combind   |              |         | 0.0000         |         | 0.0000         |         |
| Palliative * |              |         | 1.9799 (0.6336) | 0.0019  | 0.0035         |         |
| Stage grade |              |         |              |         |                |              |         |                |
| I         | 0.1649 (0.2845) | 0.5622  | 0.6460 (0.5684) | 0.2564  |
| II        | 0.8235 (0.2352) | 0.0005  | 1.2438 (0.5360) | 0.0208  | <.0001         |
| III       | 0.3421 (0.2165) | 0.1141  | <.0001         | 0.7285  | 0.5293         | 0.1695  |
| IV        | -1.9769 (0.2094) | <.0001  | -1.6439 (0.5192) | 0.0017  |
| high      | 0.7137 (0.2049) | 0.0005  | 0.9822 (0.4616) | 0.0340  |
| low*      | 0.0000         |         | 0.0000         |         |
| Spatial variable |              |         | -0.0201 (0.0082) | 0.0141  | 0.0126         | 0.0126  |
| Variance comp. |              |         |              |         |                |              |         |                |
| Sigma_{11} (\sigma_{11}^2) | /         | /        | 0.2907 (0.2339) |         |

AIC (BIC) | 1056.067 (1072.180) | 924.848 (981.244) | 923.040 (937.970) |

key: -=indicates the corresponding variable was not included in the model; /= Not applicable in the model; *=Reference categories; sr.err=standard error; comp. =component
Therefore, the final estimated model can be written by:

$$
\text{logit}(P[Y_i \leq j]) = \hat{\alpha}_j - 0.003\text{Age}_i + 0.6034\text{Anemia}_i + 1.4549\text{Chemotherapy}_i + \\
1.9799\text{Combined}_i + 0.6460\text{Stage.grade}_{Ii} + 1.2438\text{Stage.grade}_{IIIi} + \\
0.7285\text{Stage.grade}_{IIIi} - 1.6439\text{Stage.grade}_{IVi} + \\
0.9822\text{Stage.grade.high}_i - 0.0208 S_{vi} \tag{12}
$$

Since patient status have five category $J=5$, the model has four intercepts such as: $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\alpha}_3$ with its estimated value -6.2469, -1.6466, -0.03957 and 2.1793 respectively. Usually these are not of interest except for estimating the probability of patients status that fall at category $j$ or below, $P[Y\leq j]$. The spatial mixed OLR model (in Table 7) shows anemia at diagnosis, treatment, stage.grade at diagnosis and spatial variables were significant effect (p-value $\leq 0.05$) on log odds of probability of patients status category at $j$ or below. But not age. The ordered logit (log odds) for patients who hadn’t anemia complication at diagnosis being in a lower patient status category was 0.6034 greater than patients who had anemia complication at diagnosis. Hence, the estimated odds that patients without anemia complication at diagnosis is equals to 1.828 ($= e^{0.6034}$) times the estimated odds of patients who had anemia at diagnosis, which reflects that the estimated odds of patients who hadn’t anemia at diagnosis to fall in the lower direction of patient status category was higher by 82.28% of estimated odds of patients who had anemia, keeping other variables constant. This corresponds to the patients without anemia complication being less likely to fall at the higher patient status category than patients with anemia complication.

The logit (log odds) for patients who received chemotheraphy, and both chemotheraphy and surgery (combined) being in a lower patient status category were 1.4549 and 1.9799 respectively greater than patients who received only palliative care. On the other hand, the estimated odds that patients who received chemotheraphy treatment and patients who received both chemotheraphy and surgery (combined) treatment were 4.284 ($= e^{1.4549}$) and 7.242 ($= e^{1.9799}$) times the estimated odds of patients who were received palliative care respectively. Hence, for patients who were received chemotheraphy and patients who were received both chemotheraphy and surgery, the patient status to fall in the lower direction of patient status category was more likely than patients who were received palliative care. However, patients who received both chemotheraphy
and surgery were more likely to fall in the lower direction of patient status category than patients who took chemotheraphy and palliative care, Table 5.

Stage and grade (Stage grade) of patients at diagnosis was also significant effect on the log odds of probability of patients status at j or below category. The logit (log odds) of patients who had a cancer tumor at I, II, III stage and high grade being in a lower patient status category were 0.6460, 1.2438, 0.7285 and 0.9822 respectively greater than patients who had low grade of cancer tumor. However, the logit (log odds) of patients who had a cancer tumor at stage IV being in a lower patient status category was 1.6439 lower than patients who had low grade cancer tumor during diagnosis. Therefore, the estimated odds of patients who had stage I cancer tumor at diagnosis the patient status to fall in the lower direction was 1.908 (= e^{0.6460}) times the estimated odds of patients who had low grade of cancer tumor at diagnosis, which indicates that the estimated odds of patients who had stage I cancer tumor at diagnosis to fall in the lower patient status category was greater by 90.8% of the estimated odds of patients who had low grade of cancer tumor at diagnosis.

It is noted that there is spatial autocorrelation between districts. In Table 7 the p-value=0.0126 also proves that there was spatial correlation of levels of patients status between districts. The spatial variable correlation with patients status was a negative value, -0.0208, indicates that districts with lower levels of patient status are usually surrounded by districts with higher levels of patients status.

**Goodness of fit of the model**

Before fitting the data using spatial mixed OLR model first making sure whether the goodness of fit of the model was assessed using Hosmer and Lemeshow test [28]. The Hosmer and Lemeshow goodness of fit test statistic \( \chi^2 = 36.6198 \) with its corresponding p-value=0.3935 shows the model goodness of fit the data is hold.

In the fitted model (Table 7) even if the AIC (BIC) value for spatial mixed OLR model a bit less than spatial OLR model there is no any pragmatic estimated effect difference between variables and since any means of checking the proportional assumption in spatial OLR (glimmix) is not available, the proportional odds assumption of the data was checked using spatial OLR model. The score test for proportional odds assumption using chi-square shows that the chi-square
The Spatial Mixed effect Ordinal Logistic Regression Modeling of Cancer disease: West Amhara, Ethiopia

statistic 27.48 with corresponding p-value=0.061 greater than 0.05 indicating that the proportional odds assumption is satisfied. Multicollinearity was also checked using the correlation matrix (small correlation) and VIF (< 10) in which the assumption hold, i.e., suggesting no multicollinearity.

In measuring association, the values of concordant and discordant in Table 8 shows how good a model is to predict the data. Concordant value of 75.7% shows that as many as 75.7% of patients with patient status (<j) have better chance in predicting the category (<j). While the value of discordant 23.8% indicates that as many as 23.8% patients with patients status (>j) have a better chance predicting the category (<j). This shows a good implication that the size of the association in this model is very good.

Table 8: Association of predicted probabilities and observed patient status

| Concordant | Discordant |
|------------|------------|
| Percent 75.7 | 23.8 |

The Correct Classification Rate (CCR), in Table 9 of spatial mixed ordinal logistic regression model using glimmix is 54.22%. This indicates that 54.22% of patients status within the districts are predicting correctly through the model.

Table 9: CCR based on Spatial Mixed OLR model fitted using Glimmix

| Actual | Prediction | Percentage correct |
|--------|------------|--------------------|
| cured (0) | 1 | 5 | 0 | 0 | 0 | 16.67 |
| improved (1) | 0 | 178 | 2 | 0 | 11 | 93.20 |
| same (2) | 0 | 60 | 20 | 0 | 28 | 18.52 |
| deteriorate (3) | 0 | 28 | 3 | 10 | 44 | 11.76 |
| death (4) | 0 | 8 | 1 | 0 | 16 | 64.00 |
| Overall percentage correct | 54.22 |

Discussion

This paper has demonstrated the distribution of incidence of cancer disease across districts in the Western part of Amhara region, Ethiopia and determine the factors that affect patients status by considering limited number of patient characteristics from patient registry card in
FHRH. Twenty types of cancer were considered in the study and of those cancer type Breast cancer was responsible for the highest diagnosed cancer type which is also stated in the study [27] that conducted in Addis Ababa (Ethiopia) as breast cancer is the most common malignant neoplasm among women while skin, endometrial and nasophyrege were responsible for the lowest diagnosed cancer type. Among patients included in the study and received treatment only 1.45% cured and 46.02% improved while the rest are no change because of treatment and even worse after they getting treatment. The incidence of cancer within each districts were assessed. However, to make the incidence of cancer comparable between districts the population at risk was standardized [9] into the same 100,000 population at risk per district. The spatial autocorrelation, association between the value of incidence of cancer observed in a given district unit and the value of incidence of cancer observed in neighboring district units was checked using Moran’s I index and showing a negative autocorrelation.

It was found that most of patients (104 cases, 25.06% of the total) present at hospital with high stage (IV) of cancer tumor. This result also stated to the result obtained in Africa [6] and specifically in Ethiopia [29] reported that since cancer treatment is limited in the country and due to lack of awareness of the people about the disease, most cancer patients diagnosed at an advanced stage. In case of this treating the disease at advanced stage become difficult which leads to most patients have incurable from the disease and need palliative care.

Numerous studies [13][19] were reported that cancer is disease of aging , in a sense that the occurrence of cancer disease become high when peoples getting older. However, this was not case in this study that the incidence of cancer disease was predominantly concentrated on age between 41 and 55 years. It could be due to life expectancy for developed country, what the study shows in [4], is quite different from life expectancy of developing countries like Ethiopia. Mostly in Ethiopia and particularly in our study area, west Amhara peoples mostly go to religious place while they getting old instead of going to hospital to get cured from disease. In fact it might be also due to the fact that the population at risk within age group was not standardized since the population at risk within each age group wasn’t known to make it standardized. On the other hand, the age was not a significant association with patient status which indicates that the malignancy of the cancer tumor is not depends on that in which age group the patient belongs. Anemia complication of cancer disease has a significant effect on patient status that conforms
with the study in [19] and [20] which states that especially for advanced cancer anemia is a common manifestation. The estimated odds of patients who haven’t anemia at diagnosis to fall lower patient status category equals to 1.828 times the estimated odds of patients who have anemia at diagnosis, which reflects that the estimated odds of patients who haven’t anemia at diagnosis to fall lower patient status category was higher by 82.2% times of estimated odds of patients who have anemia. Patients who haven’t anemia complication at diagnosis was more likely to fall in the lower patient status category than those of patients who have anemia complication during diagnosis. It was also found that the prognostic factor stage and grade (stage_grade) during diagnosis have its own effect on the patients status after they taking appropriate treatment based on the physician prescription. It is obvious and also proved in different studies such as [13] because when grade (rate how the tumor quickly grow) and stage (extent how the tumor spread apart) increased during diagnosis the patient status will become more likely very bad since curing or improving a patient with advanced stage and grade (stage_grade) is difficult. It is advisable to have early detection of the disease so that it become simple to decreasing the malignancy of the cancer tumor in the given treatment.

ABO blood type is associated with increased risk of the lung cancer based on the study in [21]. In contrast, the study conducted in china [22] on gastric cancer explained that there is no significance association between blood group and cancer risk which is consistent with our study. From this it can be recognize that the effect of blood group on cancer risk might be depends on the cancer type.

The choice of treatment depends upon the location where the cancer tumor is located, grade, stage of the disease and health status [20] which was partially consistent with our study that the treatment type has significant effect on the patients status after they taking treatment. At the last, but not least, in this study it was found that the spatial variable has negative significant effect indicating that districts with lower levels of patient status are usually surrounded by districts with higher levels of patient status and also we explored that districts with higher incidence of cancer usually surrounded by districts with lower incidence of cancer which is in line with studies in [14][15] demonstrated that spatial epidemiology offers insight into ways that individual characteristics, community attributes and physical environments interact to produce distinctive risk, illness and disease management patterns.
Conclusion

Patients who haven’t anemia complication at diagnosis were more likely to get cure and improved based on the treatment given by physician prescription than those of patients who have anemia complication during diagnosis. Most of patients had advanced stage (IV) and grade of cancer tumor that dismantle the capability of the treatment to be less effective and increase the resistance of malignancy of the cancer tumor and hence, it was perceived that the stage and grade of patients during diagnosis have its own impact on patients status after treatment. The age of the patient when the existence of the cancer tumor disclosed haven’t any relationship with the current status of the patient after they taking treatment. If the patients have the ability to capable surgery’s treatment, on average giving both chemotherapy and surgery treatment was effective for patients to have in a good status. Districts with high incidence of cancer usually surrounded by districts with low incidence of cancer and vice versa. Similarly, it was also found that districts with patients who have severe patient status mostly surrounded by districts with patients who have improved patient status.

Declaration

Ethical Consideration

Permission to undertake the study was obtained from the ethical committees of Felege Hiwot Referral Hospital FHRH by the initiation Department of Statistics on behalf of Bahir Dar University. The study explained patients/participants who participated in the study was confidential and private information would be protected. Identification of patients/participants were done only through numerical codes and collection in the study groups were carried out only when privacy is ensured based on Ministry of health legislation in the FHRH.

Consent for publication

We, LMT and EKM, give our consent for information about ourselves and relatives to be published in Tuberculosis journal. We understand that the text and any pictures published in the article will be freely available on the Internet and may be seen by the general public. The pictures and text may also appear on other websites or in print, may be translated into other languages or used for commercial purposes. we have been offered the opportunity to read the
Availability of data and material
The datasets for generated analyses during the study was collected from Felege Hiwot Referral Hospital (FHRH), Bahir Dar, Ethiopia by personnel in the hospital by the initiation of department of statistics from Bahir Dar University.

Competing Interests
The authors declare that they have no competing interests.

Funding
No funding was obtained for this study

Authors’ Contributions
LMT had the idea of the research question for the study and proposed the first draft. LMT and EKM conducted data analysis, interpretation and prepared all figures and tables by verified the analytical method. Both authors discussed the results and contributed to the final manuscript.

Acknowledgments
we would like to thank the personnel in the hospital who helped us collect the data and expert assistance. We would also like to acknowledge our thanks, to Dr. Teshome Alene, staff of the Medical Health at college of Medicine and Health Science, Bdu, for professional assistance.

Abbreviations
FHRH Felege Hiwot Referral Hospital  CLL Chronic lymphocytic leukemia
HCC Hepatocellular carcinoma  HL Hodgkin’s lymphoma
NHL Non-Hodgkin lymphoma  RBCT Round Blue Cell Tumor
OLR Ordinary Logistic Regression  AIC Akakie Information Criterion
BIC Bayesian Information Criterion  VIF Variance Inflation Factor
GIS Geographical Information System  SOLR Spatial Ordinal Logistic Regression
SMOLR Spatial Mixed Ordinal Logistic Regression

Authors’ Details:
1,2Department of Statistics, College of Science, Bahir Dar University, P. O. Box 79 Bahir Dar, Ethiopia.
1 References

[1] Meyerson, M., Gabriel, S. and Getz, G., 2010. Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews Genetics, 11(10), p.685.

[2] National Cancer Institute at the National Institutes of Health. Retrieved May 15, 2019, from https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[3] Center, M., Siegel, R. and Jemal, A., 2015. American Cancer Society Global Cancer: Facts and Figures.

[4] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), pp.394-424.

[5] Smith, D.H., Pepin, J. and Stich, A.H., 1998. Human African trypanosomiasis: an emerging public health crisis. British medical bulletin, 54(2), pp.341-355.

[6] Jennifer, D., Cathyryne, K., 2017. Africa’s Emerging Cancer Crisis: A Call to Action

[7] De Flora, S. and La Maestra, S., 2015. Epidemiology of cancers of infectious origin and prevention strategies. Journal of preventive medicine and hygiene, 56(1), p.E15.

[8] Moan, J.E., Lagunova, Z., Bruland, Ø. and Juzeniene, A., 2010. Seasonal variations of cancer incidence and prognosis. Dermato-endocrinology, 2(2), pp.55-57.

[9] Carlin, B.P., Gelfand, A.E. and Banerjee, S., 2014. Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC.

[10] Gelfand, A.E., Diggle, P., Guttorp, P. and Fuentes, M. eds., 2010. Handbook of spatial statistics. CRC press.

[11] Waller, L.A. and Gotway, C.A.; 2004. Applied spatial statistics for public health data (Vol. 368). John Wiley and Sons.
[12] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), pp.394-424.

[13] Pakzad, R., Ghoncheh, M., Pournamdar, Z., Pakzad, I., Momenimovahed, Z., Salehiniya, H. and Makhsoesi, B.R., 2016. Spatial analysis of skin cancer incidence in Iran. Asian Pacific Journal of Cancer Prevention, 17, pp.33-37.

[14] Pakzad, R., Khani, Y., Pakzad, I., Momenimovahed, Z., Mohammadian-Hashejani, A., Salehiniya, H., Towhidi, F. and Makhsoesi, B.R., 2016. Spatial analysis of stomach cancer incidence in Iran. Asian Pacific Journal of Cancer Prevention, 17(S3), pp.27-32.

[15] Shen, X., Wang, L. and Zhu, L., 2017. Spatial analysis of regional factors and lung cancer mortality in China, 1973–2013.

[16] Brown, P., Jiang, H., Ezzat, S. and Sawka, A.M., 2016. A detailed spatial analysis on contrasting cancer incidence patterns in thyroid and lung cancer in Toronto women. BMC public health, 16(1), p.950.

[17] Roshandel, G., Boreiri, M., Sadjadi, A. and Malekzadeh, R., 2014. A diversity of cancer incidence and mortality in West Asian populations. Annals of global health, 80(5), pp.346-357.

[18] Gregorio, D.I., Cromley, E., Mrozinski, R. and Walsh, S.J., 1999. Subject loss in spatial analysis of breast cancer. Health and Place, 5(2), pp.173-177.

[19] Danaei, G., Vander Hoorn, S., Lopez, A.D., Murray, C.J., Ezzati, M. and Comparative Risk Assessment collaborating group (Cancers, 2005. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. The Lancet, 366(9499), pp.1784-1793.

[20] Benson, K., Balducci, L. and Aapro, M., 2008. Anemia and cancer. In Anemia in the Elderly (pp. 99-113). Springer, Boston, MA.

[21] Urun, Y., Utkan, G., Cangir, A.K., Oksuzoglu, O.B., Ozdemir, N., Oztuna, D.G., Kocaman, G., Coşkun, H.Ş., Kaplan, M.A., Yuksel, C. and Demirkazik, A., 2013. Association of ABO
The Spatial Mixed effect Ordinal Logistic Regression Modeling of Cancer disease: West Amhara, Ethiopia

blood group and risk of lung cancer in a multicenter study in Turkey. Asian Pac J Cancer Prev, 14(5), pp.2801-3.

[22] Xiao, S., Feng, F., Sun, L., Cai, L., Liu, Z., Liu, S., Fan, D. and Zhang, H., 2017. Blood type AB predicts promising prognosis in gastric cancer patients with positive preoperative serum CEA. Medicine, 96(47).

[23] Getahun, F., Mazengia, F., Abuhay, M. and Birhanu, Z., 2013. Comprehensive knowledge about cervical cancer is low among women in Northwest Ethiopia. BMC cancer, 13(1), p.2.

[24] Muhammad, N., Tuti Purwaningsih, S., 2013. Modeling Spatial Ordinal Logistic Regression and The Principal Component to Predict Poverty Status of Districts in Java Island, International Journal of Statistics and Applications, Vol. 3 No. 1, 2013, pp. 1-8.

[25] Muhammad, N., Tuti Purwaningsih, S., 2009. Weighted distance and optimum Intersection Spatial Logistic Regression for Prediction of Rural Poverty Status in West Java.

[26] Midi, H., Sarkar, S.K. and Rana, S., 2010. Collinearity diagnostics of binary logistic regression model. Journal of Interdisciplinary Mathematics, 13(3), pp.253-267.

[27] Hadgu, E., Seifu, D., Tigne, W., Bokretsion, Y., Bekele, A., Abebe, M., Sollie, T., Merajver, S.D., Karlsson, C. and Karlsson, M.G., 2018. Breast cancer in Ethiopia: evidence for geographic difference in the distribution of molecular subtypes in Africa. BMC women’s health, 18(1), p.40.

[28] Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. Applied logistic regression (Vol. 398). John Wiley and Sons.

[29] Woldeamanuel, Y.W., Girma, B. and Teklu, A.M., 2013. Cancer in Ethiopia. The Lancet Oncology, 14(4), pp.289-290.

[30] Boscoe, F.P., Ward, M.H. and Reynolds, P., 2004. Current practices in spatial analysis of cancer data: data characteristics and data sources for geographic studies of cancer. International journal of health geographics, 3(1), p.28.

[31] Tawhid, M., 2015. Spatial autocorrelation and its importance.
[32] Verbeke, G., and Molenberghs, G. 2000. Linear Mixed Models for Longitudinal Data. New York, NY: Springer-Verlag. Park, CA: Sage Publications.

[33] Seltman, H. (2015). Experimental design and analysis. Retrieved January 15 (2014)
Figures

Figure 1

Number of female (blue color) and male (red color) patients distribution within different cancer types
Figure 2

Incidence count of cancer disease in different age groups for female (black color) and male (red color)

Figure 3

Spatial distribution of incidence of cancer (per 100,000 population) in the respective location specified by Longitude and Latitude for each districts