Sudan University of Science and Technology

College of Graduate Studies

Determination of the Risk Factors of Prostate Cancer Incidence by Using Some Statistical Methods

A thesis submitted for the degree of doctor of philosophy in statistics.

By:

Rayyan Ibrahim Mukhtar Ahmed

Main Supervisor:

Dr. Ahmed Mohamed Abdalla Hamdi

Co- Supervisor:

Dr. Afraa Hashim Abdelatif Mohamed

Dedication

I dedicate this thesis to my parents, daughter, brothers, sister and my friends.

Acknowledgements

A sincere debt of gratitude is owed to my supervisor, Dr. Ahmed Mohamed Abdalla Hamdi, for his guidance and advice.

I would like to thank my co-supervisor, Dr. Afraa Hashim Abdelatif and the staff-members Department of Statistics, college of sciences, Sudan University of Science & Technology. Special thank for Dr. Altaiyb Omer Ahmed Mohmmed for the useful discussion.

Special thank go to my parents for their continuous encouragement and psychological & financial support.

المستخلص:

يهدف هذا البحث الى تصميم نموذج إحصائي دقيق ليصف العلاقة بين إمكانية حدوث سرطان البروستات وعوامل الخطر لهذا المرض . كما يهدف هذا البحث الي تحديد أكثر الأساليب الإحصائية ملائمة لبيانات البحث و الي تحديد عوامل الخطر للمرض التى تزيد من إنتشاره.

تم جمع البيانات من مستشفى الخرطوم للعلاج بالاشعة والطب النووي من مرضى سرطان البروستات، أخذت منهم المتغيرات التالية: العمر، المهنة، الولاية، الحالة الاجتماعية، التاريخ العائلي، تناول الدهون الحيوانية، تناول الفواكهة ، الخضروات الخضراء، زيادة الوزن ، الكوليستيرول، ضغط الدم، أدوية البروستات ، الكحول، التدخين، الاصابة بواحد أو أكثر من هذه الأمراض : الزهري، السيلان، التهاب البروستات المزمن ، تضخم البروستات ، مستضد اللبروستات المحدد. تم استخدام المنهج التحليلي لتحليل البيانات بإستخدام تحليل الإنحدار اللوجستي، إختبار كاي تربيع و إختبار مانتل هانزل لتحديد عوامل الخطر المرتبطة بحدوث المرض. لإنجاز هذا الغرض تم تحديد مرضى سرطان البروستات وأخذت عينة حجمها 250 رجلا. تم أخذ البيانات عن طريق الإستبيان ومن سجلات المرضى.

إستنتجت هذه الدراسة أن عوامل الخطر الأكثر أهمية التي إتفق عليها الطرق الإحصائية الثلاثة هى العمر و مستضد البروستات المحدد. وأيضا أظهر التحليل أن إختبار كاى تربيع هو الأفضل لتحديد عوامل الخطر لسرطان البروستات لأنه يظهر أعلى قيم كاى تربيع للمتغيرات. إستنادا على نتائج البحث هنالك بعض النقاط التى يوصى بها: الاستخدام الأمثل لإجراء مانتل هانزل في مجال الإحصاء الحيوي، نشر التوعية بين الرجال وخصوصا الذين تجاوزت أعمارهم الخمسين عاما بضرورة الفحص الدوري لمستضد البروستات المحدد، لان العمر هو عامل الخطرالأقوى لظهور سرطان البروستات.

Abstract

The aim of this research is to design a precise statistical model that shows the relation between the possibility of the incidence of prostate cancer and the disease risk factors. Also this research aims to determine the best of the three statistical methods to suit the research data related to prostate cancer and to identify the most important risk factors of the disease those increase its prevalence.

The data were collected from Khartoum Nuclear Hospital regarding the prostate cancer patients for the following independent variables were collected for cases and controls: Age, Occupation, State, Marital status, Family history, Animal fat, Fruits & Green vegetables, Overweight, Cholesterol, Blood pressure, Prostate medications, Alcohol, Smoking, Developing one or more of these diseases: 'Syphilis, gonorrhea, chronic prosatitis, and prostate enlargement', Prostate Specific Antigen (PSA). The analytical approach was used in analyzing the data by using the logistic regression analysis, chi-square test, and Mantel-Haenszel test to identify the risk factors associated with the occurrence of the disease. In order to implement this, prostate cancer patients were specified, and sample of 250 men was taken. The data were collected through a questionnaire, and from the patient records.

This study concluded that the most important risk factors that agreed by all three procedures were age and PSA. The analysis also showed that chi-square test is the best in terms of determining the risk factors for the disease because it contains the highest χ^2 values for the variables. Based on the research findings the following points are to be recommended: Maximum use of the Mantel-Haenszel procedure in the biostatistics field, Raising awareness of the need to examine PSA periodically, especially when the age equal to or above 50 years, because age is the strongest risk factor for the appearance of prostate cancer.

Table of Contents:

List of Tables:

Chapter 1

Introduction

Chapter one

(1.1) Preface:

Since ancient times, many diseases that used to affect the human body occurred which can be diagnosed and treated early so that the person can survive and be productive. Also, there are some diseases that are symptomless and their signs would be felt by the patient only in advanced phases of the disease. In such cases, treatment would be complicated and the body's response for treatment would be low, unlike the treatment at earlier stages of the disease. Such diseases are like viral hepatitis, heart diseases, diabetes, and different types of cancer.

Cancer is considered one of the difficult challenges that face health authorities due to its high prevalence and increasing number of cases of the disease. It claims the lives of most of the people who are diagnosed with it. In the USA, one out of two persons is diagnosed with a type of cancer in his lifetime. It claims the lives of 22% of people, putting it as the second cause of death in advanced countries; it is expected to be the main cause of death in the coming years. Cancer has many negative impacts on the patient and the society as a whole, the word "cancer" is an alarming word for its patient, and he is prone to many psychological, social and financial effects. So, by early diagnosis, the disease would be treatable, the thing which would be a turning point in the life of the patient and his family.

Cancer affects different age categories, but it is more obvious with advanced age, and most types of cancer affect people after their 50s. Prostate cancer is an important type of cancer and claims so many lives in advanced countries, and it is more in developing countries. The Regional Conference on Oncology and Hematology provided the latest statistics regarding the prevalence of prostate cancer in the Middle East and North Africa. The figures are five times less than that in the US, but the death rate is five times more than that in the USA.

In Sudan, prostate cancer ranks second in the statistics of the Health

Ministry- Sudan, (Sudanese Ministry of Health 2016), but the reasons for this type of cancer are not identified. It is very important to conduct a study to identify the risk factors that cause this type of cancer by using some statistical methods such as the Logistic Regression, chi-square test and Mantel-Haenszel test in order to reduce the incidence of the disease.

(1.2) Research Problem

The age categories that are more affected by prostate cancer are men above 50, but this does not mean that other age categories are cancer-free. It is considered a silent killer, because no signs or symptoms are noticed on the patient in the early stages of the disease. Also, scientists could not figure out its causes. It is important to conduct a detailed study on prostate cancer, its symptoms, diagnosis, and risk factors, so that it would be reduced.

(1.3) The Importance of The Research :

A number of researches concentrated on cancer patients by using different statistical methods. All those researches reached different conclusions that aimed at reducing the prevalence of the disease in various age categories. This study tackles prostate cancer by building an accurate statistical model to determine the most factors that related to the disease and to predict a person's risk of prostate cancer by using logistic regression model. Also Chi-square test (sometimes Fisher exact test) and Mantel-Haenszel test are used to select the risk factors which contribute to the appearance of the disease, which makes this research distinguished from other studies by highlighting the use of these methods in medical and epidemiological researches.

(1.4) The Objectives of The Research:

1. To design a precise statistical model that shows the relation between the possibility of the incidence of prostate cancer (dependent variable) and the disease risk factors (independent variables).

2. To determine the best of the two statistical tests to suit the research data related to cancer.

3. To identify the most important risk factors of the disease those increase its prevalence.

4. To assess the parameters of statistical model for data representation

(1.5) Research Hypothesis:

1. There is a clear preference when using chi-square test and Mantel–Haenszel test to determine the risk factors.
2. The following independent variables have effect on the incidence of prostate cancer, they are: Age, Occupation, State , Marital status, Family history, Animal fat, Fruits and green vegetables, Overweight, Cholesterol, Blood pressure, Prostate medications, Alcohol, Smoking, Developing one or more of these diseases: 'Syphilis, gonorrhea, chronic prosatitis, and prostate enlargement' and Prostate Specific Antigen (PSA).
3. The logistic regression model has ability to calculate the probability of prostate cancer incidence when the values of the risk factors are known.

(1.6) Research Limitations:

The study was limited to the prostate cancer patients who get medical treatment at Khartoum Nuclear Hospital, the researcher used to visit this hospital on a daily basis in the period between 24-4-2016 and 20-12-2016.

(1.7) Research Data:

The data were collected from Khartoum Nuclear Hospital " Amal Tower" regarding the prostate cancer patients for the following independent variables were collected for cases and controls: Age, Occupation, State , Marital status, Family history, Animal fat, Fruits & Green vegetables, Overweight, Cholesterol, Blood pressure, Prostate medications, Alcohol, Smoking, Developing one or more of these diseases: 'Syphilis, gonorrhea, chronic prosatitis, and prostate enlargement' and Prostate Specific Antigen (PSA).

(1.8) Research Methodology:

The researcher used the descriptive approach to describe the study variables for prostate cancer patients in all stages of the disease. Also, the analytical approach was used in analyzing the data by using the logistic regression analysis, chi-square test (sometimes fisher test), and Mantel-Haenszel test to identify the risk factors associated with the occurrence of the disease. A statistically significant model was also conducted to describe the relationship between the dependent variable and the independent variables that represent the risk factors of the disease accurately so that it is predicted in the future. In order to implement this, prostate cancer patients were specified, and sample of 250 men was taken. The data were collected through a questionnaire, and from the patient records.

(1.9) Research Structure:

This study contains five chapters, chapter one discussed the research problem, importance, objectives, hypothesis, limitations, data, and methodology. It also included a number of previous studies. Chapter two dealt with identifying prostate cancer, its risk factors, symptoms, diagnosis, stages, treatment methods, and some patient statistics. As for chapter three, it was concerned with explaining the statistical method used in the study, it included the Logistic Regression model, chi-square test, and Mantel-Heanszel test. Chapter four focused on the practical side of the research, the study data were discussed, and the hypothesis of the variables and their validity were tested .Chapter five, the concluding chapter, showed the study results, and the recommendations that would be of benefit for the concerned people.

(1.10) Review Studies:

1) AB Weiner et al. conducted a study in 2017 published in journal of Prostate Cancer and Prostatic Diseases in title "Contemporary management of men with high-risk localized prostate cancer in the United States". Using the National Cancer Data Base for 2004–2013, all men diagnosed with high-risk localized prostate cancer (PCa) were identified using National Comprehensive Cancer Network criteria. Temporal trends in

initial management were assessed. Multivariable logistic regression was used to evaluate demographic and clinical factors associated with undergoing radical prostatectomy (RP). In total, 127 391 men were identified. Use of RP increased from 26% in 2004 to 42% in 2013 (adjusted risk ratio (RR) 1.51, 95% CI: 1.42–1.60, P<0.001), while external beam radiation therapy (EBRT) decreased from 49% to 42% (P<0.001). African American men had lower odds of undergoing RP (unadjusted rate of 28%, adjusted RR 0.69, 95% CI: 0.66–0.72, P<0.001) compared to White men (37%). Age was inversely associated with likelihood of receiving RP. Having private insurance was significantly associated with the increased use of RP (vs. Medicare, adjusted odds ratio 1.04, 95% CI: 1.01–1.08, P = 0.015). Biopsy Gleason scores 8–10 with and without any primary Gleason 5 pattern were associated with decreased odds of RP (vs. Gleason score ≤ 6, both P<0.001). Academic and comprehensive cancer centers were more likely to perform RP compared to community hospitals (both P<0.001). The researchers concluded that the likelihood of receiving RP for high-risk prostate cancer dramatically increased from 2004 to 2013. By 2013, the use of RP and EBRT were similar. African American men, elderly men and those without private insurance were less likely to receive RP. (Weiner, A.B. et al. 2017).

2) Another study was conducted by M.I. Gökce et al. in 2017 in title "Is active surveillance a suitable option for African American men with prostate cancer? A Systemic Literature Review". This paper was published in journal of Prostate Cancer and Prostatic Diseases. A literature review through the Medline database published from 1990 until August 2015 was performed to identify studies reporting outcomes of the African American (AA) population with low-risk prostate cancer that underwent either active surveillance (AS) or treatment. An additional search for studies on genetic mechanisms involved in development of prostate cancer in AA men was also performed. Eleven studies on pathologic results of AA men who would qualify for AS were identified and in eight of these studies AA race was found to be associated with adverse pathological outcomes such as positive surgical margins, upgrading or upstaging. The other three studies reported no significance in these parameters with respect to race. Five more studies

reported outcomes of AS in AA men with different study end points. AA men were mainly found to have a higher rate of disease reclassification subsequent to active treatment. The studies on genetic mechanisms also identified different genetic alterations in the AA population. Thus this study concluded that the AA men with clinically defined low-risk prostate cancer may have either a higher grade or volume of cancer that was not detected on routine evaluation. Therefore, AS among such patients should be approached with caution. The researchers recommended discussing such risks with AA patients with an acknowledgement that existing favorable outcomes noted in largely Caucasian populations may not be applicable to AA patients. They proposed a modified evaluation plan for AA patients that included an early confirmatory biopsy preceded by a magnetic resonance imaging to optimally detect occult cancer foci, (Gökce, M.I. et al. 2017).

3) DM .Moreira et al. introduced a scientific paper in 2017 with title "The combination of histological prostate atrophy and inflammation is associated with lower risk of prostate cancer in biopsy specimens" which published in journal of Prostate Cancer and Prostatic Diseases. To evaluate whether the presence of both prostate atrophy (PA) and chronic prostate inflammation (CPI) in the same biopsy and in the same biopsy core are associated with prostate cancer risk and grade in repeat biopsies, a retrospective analyses of 6132 men who were 50–75 years old undergoing 2-year repeat prostate biopsy after a negative baseline biopsy for prostate cancer (PCa) in the REduction by DUtasteride of prostate Cancer Events (REDUCE) study. Prostate atrophy (PA), chronic prostate inflammation (CPI) and PCa were determined by central pathology. The association of baseline PA and CPI with 2-year repeat biopsy cancer status and grade was evaluated with χ^2 test and logistic regression controlling clinicopathological features. PA, CPI and both were detected in 583 (9.5%), 1063 (17.4%) and 3675 (59.9%) baseline biopsies, respectively. Compared with biopsies with neither PA nor CPI, the presence of PA (odds ratio (OR) = 0.73, 95% confidence interval (CI) = 0.57–0.93), CPI (OR = 0.72, 95% CI = 0.58–0.88) and both (OR = 0.54, 95% CI = 0.45–0.64) were associated with lower PCa risk in the 2-year repeat prostate biopsy. Results were similar in

multivariable analysis. Among subjects with both PA and CPI, those with both findings in the same core had even lower PCa risk compared with PA and CPI in different cores (univariable OR = 0.68, 95% CI = 0.51–0.91; multivariable OR = 0.73, 95% CI = 0.54–0.99). Combination of PA and CPI was associated with lower risk of high-grade PCa. They concluded that the presence of both PA and CPI in baseline biopsies, especially in the same core, was associated with lower PCa risk and grade. The presence and topographical distribution of PA and CPI may be used in PCa risk stratification, (Moreira, D.M .et al. 2017).

4) AC .Vidal and LE. Howard et al. conducted analytical study in 2017 which published in journal of Prostate Cancer and Prostatic Diseases with title "Obesity and prostate cancer-specific mortality after radical prostatectomy: results from the Shared Equal Access Regional Cancer Hospital (SEARCH) database". They conducted a retrospective analysis of 4268 radical prostatectomy patients within the Shared Equal Access Regional Cancer Hospital (SEARCH) database. Cox models accounting for known risk factors were used to examine the associations between body mass index (BMI) and PC-specific mortality (PCSM; primary outcome). Secondary outcomes included biochemical recurrence (BCR) castration-resistant PC (CRPC). BMI was used as a continuous and categorical variable (normal < 25 kg/m^2, overweight 25–29.9 kg/m^2 and obese ≥ 30 kg/m^2). Median follow-up among all men who were alive at last follow-up was 6.8 years. During this time, 1384 men developed BCR, 117 developed CRPC and 84 died from PC. Hazard ratios were analyzed using competing-risks regression analysis accounting for non-PC death as a competing risk. On crude analysis, higher BMI was not associated with risk of PCSM (P = 0.112), BCR (0.259) and CRPC (P = 0.277). However, when BMI was categorized, overweight (hazard ratio (HR) 1.99, P = 0.034) and obesity (HR 1.97, P = 0.048) were significantly associated with PCSM. Obesity and overweight were not associated with BCR or CRPC (all P $\geq$ 0.189). On multivariable analysis adjusting for both clinical and pathological features, results were little changed in that obesity (HR = 2.05, P = 0.039) and overweight (HR = 1.88, P = 0.061) were associated with higher risk of PCSM, but not with BCR or CRPC (all P $\geq$ 0.114) with the exception that the association

for overweight was no longer statistical significant. They concluded that overweight and obesity were associated with increased risk of PCSM after radical prostatectomy. They suggested that if validated in larger studies with longer follow-up, obesity may be established as a potentially modifiable risk factor for PCSM, (Vidal, A.C. and Howard, L.E. et al.).

5) M. Gacci and G. I. Russo et al. conducted analytical study in 2017 which published in Prostate Cancer and Prostatic Diseases, in title "Meta-analysis of metabolic syndrome and prostate cancer". The aims of this study were to evaluate the impact of metabolic syndrome and metabolic syndrome factors on Prostate cancer incidence, on the risk of high-grade Prostate cancer and to analyze the role of metabolic syndrome and single metabolic syndrome components on the development of aggressive Prostate cancer features, (Gacci, M. and Russo, G. I. et al. 2017).

6) Paul R. Rosenbaum and Dylan S. Small introduced a paper in 2016 which publised in Journal of The International Biometric and Society with title "An Adaptive Mantel–Haenszel Test for Sensitivity Analysis in Observational Studies'. They proposed a sensitivity analysis for an adaptive test similar to the Mantel–Haenszel test. The adaptive test performed two highly correlated analyses, one focused analysis using a subgroup, one combined analysis using all of the data, correcting for multiple testing using the joint distribution of the two test statistics. Because the two component tests were highly correlated, this correction for multiple testing was small compared with, for instance, the Bonferroni inequality. The test had the maximum design sensitivity of two component tests. A simulation evaluated the power of a sensitivity analysis using the adaptive test. Two examples were presented. An R package, sensitivity2x2xk, implemented the procedure, (Rosenbaum, Paul R. and Dylan, S. 2016).

7) Davies Adeloye and Rotimi Adedeji David et al. introduced a scientific paper in 2016 which published in PLOS One Journal in title "An Estimate of the Incidence of Prostate Cancer in Africa: A Systematic Review and Meta-Analysis". The researchers systematically reviewed the literature on

prostate cancer in Africa and provided a continent wide incidence rate of prostate cancer based on available data in the region and they conducted a random effects meta-analysis. Their search returned 9766 records, with 40 studies spreading across 16 African countries meeting their selection criteria. We estimated a pooled prostate cancer incidence rate of 22.0 (95% CI: 19.93–23.97) per 100,000 population, and also reported a median incidence rate of 19.5 per 100,000 population. They observed an increasing trend in prostate cancer incidence with advancing age, and over the main years covered, (Adeloye, Davies and David, Rotimi Adedeji et al. 2016).

8) L. Nicholson and H. Hotchin introduced a scientific paper in 2015 in Journal of Intellectual Disability Research. Its title was" The relationship between area deprivation and contact with community intellectual disability psychiatry". This study investigated the relationship between area deprivation and contact with intellectual disabilities psychiatry. Psychiatric case notes and electronic records were used to identify all patients who had face-to-face contact with community intellectual disabilities (ID) psychiatric services over 1 year in the North East Community Health Partnership of Greater Glasgow and Clyde (estimated population 177 867). The Scottish Index of Multiple Deprivation (SIMD) was determined for the patient sample (553 patients) and for the general population living in the same area. IBM SPSS statistics version 19 was used to analyze the data. Descriptive statistics were used to describe the sample and general population SIMD data and they were compared using Fisher's Exact and Mann–Whitney U tests. They concluded that in the area under study, contact with ID psychiatry was greater in more deprived areas. Given the high psychiatric morbidity of people with ID, if services do not adjust for deprivation, this may lead to further discrimination in an already disadvantaged population (Nicholson, L. and Hotchin, H. 2015).
9) M. Moosazadeh et al. in 2015 conducted analytical study in Eastern Mediterranean Health Journals, in title "Predictive factors of death in patients with tuberculosis: a nested case–control study" , This study aimed to determine predictive factors for death in patients with tuberculosis to set priorities

for public heath interventions to reduce mortality in these patients. This nested case–control study was carried out in Mazandaran province of Islamic Republic of Iran among tuberculosis patients who were treated during 2002–2009. Each deceased patient was individually matched with a control patient according to sex, age, area of involvement and time of follow-up. Potential risk factors for death were evaluated using multivariate conditional logistic regression models. From 2206 patients 376 cases and 376 matched controls were selected. Only positive serology for HIV (OR = 19.1), history of kidney disease (OR = 6.81) and use of immunosuppressant drugs (OR = 3.96) significantly increased the risk of death in tuberculosis patients. These potentially modifiable risk factors could be taken into account in preventive interventions for tuberculosis patients in our country, (Moosazadeh, M. et al. 2015).

10) Mathias Barra et al. conducted study in 2014 which published in Journal of Head and Face Pain, with title "Statistical Testing of Association between Menstruation and Migraine". The objective of this study was to repair and refine a previously proposed method for statistical analysis of association between migraine and menstruation. The statistical method was based on a simple two-parameter null model of the menstrual related migraine MRM (which allows for simulation modeling), and Fisher's exact test (with mid-p correction) applied to standard 2 × 2 contingency tables derived from the patients' headache diaries. Their method was a corrected version of a previously published flawed framework. In this paper, they corrected a proposed method for establishing association between menstruation and migraine by statistical methods. They concluded that the proposed standard of 3-cycle observations prior to setting an MRM diagnosis should be extended with at least one premenstrual window to obtain sufficient information for statistical processing, (Barra, Mathias et al. 2014).

11) Vaclav Fidler and Nico Nagelkerke conducted a paper in 2013 which published in PLOS ONE Journal, with title "The Mantel-Haenszel Procedure Revisited: Models and Generalizations". Here we revisit the Mantel-Haenszel method and propose an extension to continuous and vector valued Z. The idea is to replace the observed cell entries in strata of the

Mantel-Haenszel procedure by subject specific classification probabilities for the four possible values of (X, Y) predicted by a suitable statistical model. For situations where X and Y can be treated symmetrically we propose and explore the multinomial logistic model. Under the homogeneity hypothesis, which states that the odds ratio does not depend on some confounder Z, the logarithm of the odds ratio estimator can be expressed as a simple linear combination of three parameters of this model. Methods for testing the homogeneity hypothesis are proposed. The relationship between this method and binary logistic regression is explored. A numerical example using survey data is presented, (Fidler, Vaclav and Nagelkerke, Nico. 2013).

12) Another paper was introduced by Adejumo, A. O. and Adetunji, A. A. in 2013 with title "Cochran–Mantel–Haenszel Test for Repeated Tests of Independence: An Application in Examining Students' Performance", which accepted and published in Journal of Education and Practice. The researchers were interested in collecting information for each of several 2 x 2 tables across the levels of the subpopulations. From the result of graduate of ten departments in Faculty of Science, University of Ilorin for 2011/2012 academic session, data on final cumulative grade point average (Final Grade); department (ten departments of the faculty); age at entry (below or 20 years and above 20 years) and sex (male and female) were analyzed using Cochran-Mantel-Haenszel statistics. Odds of a student graduating with second class upper and above (0.5270) was about half of graduating with second class lower and below. This implied that the final grade was approximately symmetrical about two groups. The students in first group were those with second class lower and below (Low Grade) while the other was for those with second class upper and above (High Grade). Breslow-Day and Tarone's statistics showed that the null hypothesis of homogeneity of odds ratio across the departments was not rejected for both age at entry and sex. This implied that the odds ratio across the ten departments (relating to age at entry & final grade and sex & final grade) were all equal. Cochran's and Mantel-Haenszel statistics revealed the final grade of students (Low Grade or High Grade) was not associated with both sex and age students at entry. The odds in

favour of a student whose age was less than 20 years graduating with Low Grade (Pass, Third Class, and Second Class Lower) was 0.865 while it was 0.670 for male students graduating with lower grade, (Adejumo, A. O. and Adetunji, A. A. 2013).

13) Yuko Kanbayashi et al. introduced a scientific paper in 2013 with title "Predictive Factors for Agitation Severity of Hyperactive Delirium in Terminally Ill Cancer Patients in a General Hospital Using Ordered Logistic Regression Analysis" which published in journal of Palliative Medicine. This study aimed to identify predictive factors for agitation severity of hyperactive delirium in terminally ill cancer patients in a general hospital. Participants were 182 consecutively admitted terminally ill cancer patients who died in a Japanese general hospital between April 2009 and March 2011. Variables present one week before death were extracted from the clinical records for regression analysis of factors potentially related to agitation severity of delirium. The prevalence and agitation severity of delirium were evaluated retrospectively. Multivariate ordered logistic regression analysis was performed to identify predictive factors. Male sex [odds ratio (OR) = 2.125, 95% confidence interval (CI) = 1.111–4.067; P = 0.0227]; total bilirubin (T-bil) [OR = 1.557, CI = 1.082–2.239; P = 0.017]; antibiotics [OR = 0.450, CI = 0.219–0.925; P = 0.0298]; nonsteroidal antiinflammatory drugs (NSAIDs) [OR = 2.608, CI = 1.374–4.950; P = 0.0034]; and hematological malignancy [OR = 3.903, CI = 1.363–11.179; P = 0.0112] were found to be statistically significant predictors for agitation severity of hyperactive delirium. Their study indicates that male sex, T-bil, antibiotic therapy, NSAID therapy, and hematological malignancy are significant predictors for agitation severity of hyperactive delirium in terminally ill cancer patients in a general hospital setting, (Kanbayashi, Yuko et al. 2013).

14) Sureiman Onchiri conducted study in 2013 which published in Educational Research and Reviews with title "Conceptual model on application of chi-square test in education and social sciences". Chi-square test is one of the most frequently used tests with a number of improper applications. Some of the general causes of the improper applications include researchers not understanding the areas and conditions of application of the Chi square test. To give

solutions to the above problems, this paper explored the existing literature on the main areas of application of Chi-square, that include the test of frequencies (goodness of fit, homogeneity, independence) and the test of population variance. The paper identifies the shortfalls in the existing literature, and fills them by the application of appropriate illustrations and examples. To shield the loopholes in data analysis using Chi-square test, a simplified conceptual model which can be adopted by researchers is finally developed, (Onchiri, Sureiman. 2013).

15) John Ludbrook in 3013 introduced a review in journal of the Clinical and Experimental Pharmacology and Physiology with title" Analyzing 2x2 contingency tables: Which test is best?" He summarized that; a survey of five journals of physiology or pharmacology for 2011 showed that Fisher's exact test was used three times as frequently as Pearson's Chi-squared test, Pearson's test requires that random samples are taken from defined populations. The resultant 2x2 table is described as unconditional because neither the row nor column marginal totals are fixed in advance. Also he noticed that Fisher's test requires the rare condition that both row and column marginal totals are fixed in advance. The resultant 2x2 table is described as doubly conditioned. However, the most common design of biomedical studies is that a sample of convenience is taken and divided randomly into two groups of predetermined size. The groups are then exposed to different sets of conditions. The binomial outcome is not fixed in advance, but depends on the result of the study. Thus, only the column (group) marginal totals are fixed in advance and the table is described as singly conditioned. Singly conditioned 2x2 tables are best analyzed by tests of null hypotheses on the odds ratio (OR= 1) or by tests on proportions (p), such as the relative risk (RR=p2/p1=1) or the difference between proportions (p2–p1 = 0). One enormous advantage of these procedures is that they test specific hypotheses. They should be executed in an exact fashion by permutation, (Ludbrook, John. 3013).

16) B.Zhang et al conducted a study in 2011; it was published in The Journal of International Medical Research with title "Assessment of Risk Factors for Early Seizures Following Surgery for Meningiomas Using Logistic Regression Analysis". This study analyzed the influence of clinical factors on early postoperative seizures in patients with meningiomas and constructed a logistic regression equation for assessing risk factor. Clinical data from 222 patients with meningiomas were collected. The odds ratios (ORs) for independent variables were determined: the ORs for preoperative seizure history and movement disorder were > 1, whereas the OR for prophylactic therapy was < 1. Logistic regression analysis was then performed to select potential risk factors for early postoperative seizures. Five variables (preoperative seizure history, movement disorder, tumor location, primary location of initial tumor and prophylactic therapy) were introduced into the regression model. A logistic regression equation was then constructed that had a positive predictive value of 66.65% and a negative predictive value of 84.95%. This suggested that the five variables introduced in the equation were closely associated with early postoperative seizures, with preoperative seizure history and movement disorder as potential risk factors and prophylactic therapy as a protective factor, (Zhang, B. et al. 2011).

17) Sorana D. Bolboacă et al. introduced paper in 2011 which published in the Information Journal in title "Pearson-Fisher Chi-Square Statistic Revisited". The aim of this paper was to present solutions to common problems when applying the Chi-square tests for testing goodness-of-fit, homogeneity and independence. The main problems identified in the application of the goodness-of-fit test were as follows: defining the frequency classes, calculating the X^2 statistic, and applying the χ^2 test. Several solutions were identified, presented and analyzed. Three different equations were identified as being able to determine the contribution of each factor on three hypothesizes (minimization of variance, minimization of square coefficient of variation and minimization of X^2 statistic) in the application of the Chi-square test of homogeneity. The best solution was directly related to the distribution of the experimental error. The Fisher exact test proved to be the

"golden test" in analyzing the independence while the Yates and Mantel-Haenszel corrections could be applied as alternative tests, (Bolboacă, Sorana D. et al. 2011).

18) Another paper introduced by Todd Michael Frank et al. in 2011. It published in American Journal of Evaluation, with title "the chi-square test Often Used and More Often Misinterpreted". This paper attempts to clarify any confusion about the uses and interpretations of the family of chi-square tests developed by Pearson, focusing primarily on the chi-square tests of independence and homogeneity of variance (identity of distributions). A brief survey of the recent evaluation literature is presented to illustrate the prevalence of the chi-square test and to offer examples of how these tests are misinterpreted. While the omnibus form of all three tests in the Karl Pearson family of chi-square tests—independence, homogeneity, and goodness-of-fit,—use essentially the same formula, each of these three tests is, in fact, distinct with specific hypotheses, sampling approaches, interpretations, and options following rejection of the null hypothesis. Finally, a little known option, the use and interpretation of post hoc comparisons based on Goodman's procedure following the rejection of the chi-square test of homogeneity, is described in detail, (Frank, Todd Michael et al. 2011).

19) Giovanni Tripepi et al. conducted a study in 2010 in Nephron Clinical Practice Journal in title "Stratification for Confounding –Part1: The Mantel-Haenszel Formula". The Mantel-Haenszel formula was applied in cohort and in case control studies to calculate an overall, un compounded, effect estimate of a given exposure for a specific outcome by combining stratum-specific relative risks (RR) or odds ratios (OR). Stratum-specific RRs or ORs were calculated within each stratum of the confounding variable and compared with the corresponding effect estimates in the whole group (that was, with the un stratified RR or OR). The use of the Mantel-Haenszel formula presented some limitations: (1) if there was more than a single confounder, the application of this formula was laborious and demands a relatively large sample size, and (2) this method requires continuous confounders to be constrained into a limited number of categories thus potentially generating residual confounding (a phenomenon particularly

relevant when the variable is categorized into few strata). In the stratifSied analysis, residual confounding can be minimized by increasing the number of strata, a possibility strictly dependent on sample size, (Tripepi, Giovanni et al. 2010).

20) Another study conducted by the researchers Chao-Ying Joanne Peng et al in (2002) in title "An Introduction to Logistic Regression Analysis and Reporting". This study had been published in journal of educational research. The purpose of this article is to provide researchers, editors, and readers with a set of guidelines for what to expect in an article using logistic regression techniques. They showed that the preferred pattern for the application of logistic methods with an illustration of logistic regression applied to a data set in testing a research hypothesis. They demonstrated that logistic regression can be a powerful analytical technique for use when the outcome variable is dichotomous. The effectiveness of the logistic model was shown to be supported by (a) significance tests of the model against the null model, (b) the significance test of each predictor, (c) and predicted probabilities, (Peng, Chao-Ying Joanne et al. 2002).

This study has distinguished on the previous studies, because it combined three statistical methods; the Mantle-Haenszel test, the chi square test, and the logistic regression, and comparisons between the two tests were conducted. This only study that identified 15 variables to determine the most important variables that affect on the incidence of prostate cancer.

Chapter 2
Literature Review

Chapter Two

(2.1) Preface:

Since ancient times, many diseases that used to affect the human body occurred which can be diagnosed and treated early so that the person can survive and be productive. Also, there are some diseases that are symptomless and their signs would be felt by the patient only in advanced phases of the disease. In such cases, treatment would be complicated and the body's response for treatment would be low, unlike the treatment at earlier stages of the disease. Such diseases are like viral hepatitis, heart diseases, diabetes, and different types of cancer.

Cancer is considered one of the difficult challenges that face health authorities due to its high prevalence and increasing number of cases of the disease. It claims the lives of most of the people who are diagnosed with it. In the US, one out of two persons is diagnosed with a type of cancer in his lifetime. It claims the lives of 22% of people, putting it as the second cause of death in advanced countries; it is expected to be the main cause of death in the coming years. Every year, the American Cancer Society estimates the numbers of new cancer cases and deaths that will occur in the United States, in the current year and compiles the most recent data on cancer incidence, mortality, and survival. The latest cancer statistics estimates new cases and deaths are 180,890 and 26,120 respectively, (American Cancer Society 2016[a]).

Prostate cancer is a deadly disease that affects millions of men each year. There are few people who have not been touched personally or through friends or family by this disease. Before the twentieth century it was hardly recognized as a disease by doctors. A number of years ago, hundreds of thousands of men will be diagnosed with prostate cancer—unfortunately it happens every year. While prostate cancer is considered a disease of old men, it now can be found in middle-aged men, (Cramer, Scott D. 2007).

Cancer has many negative impacts on the patient and the society as a whole, the word "cancer" is an alarming word for its patient, and he is prone

to many psychological, social and financial effects. So, by early diagnosis, the disease would be treatable, the thing which would be a turning point in the life of the patient and his family.

Cancer affects different age categories, but it is more obvious with advanced age, and most types of cancer affect people after their 50s. Prostate cancer is an important type of cancer and claims so many lives in advanced countries, and it is more in developing countries. In Sudan, prostate cancer ranks second in the statistics of the Health Ministry- Sudan- in the year (2016), 33.1% of men were affected with prostate cancer, (Report of the Sudanese Federal Ministry of Health. 2016).

(2.2) Definition of the Prostate Cancer:

Only men have a prostate. It is a small gland that sits below the bladder near the rectum. It surrounds the urethra, the passage in the penis through which urine and semen pass. The prostate gland is part of the male reproductive system. It produces most of the fluid that makes up semen that enriches sperm. The prostate needs the male hormone testosterone to grow and develop. The prostate is often described as being the size of a walnut and it is normal for it to grow as men age. Sometimes this can cause problems, such as difficulty urinating. These problems are common in older men and not always symptoms or signs of cancer.

Prostate cancer occurs when abnormal cells develop in the prostate. These abnormal cells can continue to multiply in an uncontrolled way and sometimes spread outside the prostate into nearby or distant parts of the body. Prostate cancer is generally a slow growing disease and the majority of men with low grade prostate cancer live for many years without symptoms and without it spreading and becoming life-threatening. However, high grade disease spreads quickly and can be lethal. Appropriate management is the key, (Prostate Cancer Foundation of Australia 2016).

(2.3) Key Statistics for Prostate Cancer:

Other than skin cancer, prostate cancer is the most common cancer in American men. The American Cancer Society's estimates for prostate

cancer in the United States for 2017 are:

About 161,360 are new cases of prostate cancer and 26,730 deaths from prostate cancer. Prostate cancer is the third leading cause of cancer death in American men, behind lung cancer and colorectal cancer. About 1 man in 39 will die of prostate cancer, (American Cancer Society 2016[b]). In Africa the numbers refers to 29,530 (31%) of men were affected with prostate cancer in 2016 and about 4,450 (12%) deaths, (American Cancer Society 2016[c]).

(2.4) Causes of the Prostate Cancer:

The exact causes of prostate cancer are not known. Several risk factors for developing prostate cancer have been identified, but which of these risk factors cause a prostate cell to become cancerous is not fully known. For a cancer to develop, changes must occur in the chemicals that make up the DNA, which makes up the genes in the cell. The genes control how the cell works, for instance, how quickly the cell grows, divides into new cells, and dies, as well as correcting any mistakes that occur in the DNA of the cell to keep the cell working normally. Cancer occurs when certain genes that either control the growth or death of the cell are affected, which results in abnormal cell growth and/or death. Genes are inherited, so it passed on from parents to their children, and thus some changes in the genes (gene mutations) that increase the risk of developing cancer may be inherited. For prostate cancer, approximately 5%-10% of prostate cancers are due to inherited gene changes. Gene changes may also be acquired. These changes are not passed on to children. Such changes may occur when a cell is normally undergoing growth and division. It is thought that at times during normal cell growth, risk factors may affect the DNA of the cell, (Pamela I. Ellsworth 2016).

Some genes control when our cells grow, divide into new cells, and die:

- Certain genes that help cells grow, divide, and stay alive are called oncogenes.
- Genes that normally keep cell growth under control, repair mistakes in DNA, or cause cells to die at the right time are called tumor suppressor genes. Cancer can be caused in part by DNA changes that turn on oncogenes or turn off tumor suppressor genes.

DNA changes can be:

I. Inherited Gene Mutations:

Cancer caused by inherited genes is called hereditary cancer. Several inherited mutated genes have been linked to hereditary prostate cancer, including:

RNASEL (formerly HPC1): The normal function of this tumor suppressor gene is to help cells die when something goes wrong inside them. Inherited mutations in this gene might let abnormal cells live longer than they should, which can lead to an increased risk of prostate cancer.

BRCA1 and BRCA2: These tumor suppressor genes normally help repair mistakes in a cell's DNA (or cause the cell to die if the mistake can't be fixed). Inherited mutations in these genes more commonly cause breast and ovarian cancer in women. But changes in these genes (especially *BRCA2*) also account for a small number of prostate cancers.

DNA mismatch repair genes (such as MSH2 and MLH1): These genes normally help fix mistakes (mismatches) in DNA that are made when a cell is preparing to divide into 2 new cells. (Cells must make a new copy of their DNA each time they divide.) The inherited mutations in these genes have), increased risk of colorectal, prostate, and some other cancers.

HOXB13: This gene is important in the development of the prostate gland. Mutations in this gene have been linked to early-onset prostate cancer (prostate cancer diagnosed at a young age) that runs in some families. Fortunately, this mutation is rare.

II. Acquired Gene Mutations:

Some gene mutations happen during a person's lifetime and are not passed on to children. These changes are found only in cells that come from the original mutated cell. These are called acquired mutations. Most gene mutations related to prostate cancer seem to develop during a man's life rather than having been inherited. Every time a cell prepares to divide into 2 new cells, it must copy its DNA. This process is not perfect, and sometimes errors occur, leaving defective DNA in the new cell. It's not clear how often these DNA changes might be random events, and how often they are influenced by other factors. In general, the more quickly prostate cells grow and divide the more chances there are for mutations to occur. Therefore, anything that speeds up this process may make prostate cancer more likely, (American Cancer Society 2016[d]).

(2.5) Prostate Cancer Risk Factors:

A risk factor is anything that affects your chance of getting a disease such as cancer. Different cancers have different risk factors. Some risk

factors, like smoking, can be changed. Others, like a person's age or family history, can't be changed. But having a risk factor, or even several, does not mean that a person will get the disease. Many people with one or more risk factors never get cancer, while others who get cancer may have had few or no known risk factors. Researchers have found several factors that might affect a man's risk of getting prostate cancer:

Age: Prostate cancer is rare in men younger than 40, but the chance of having prostate cancer rises rapidly after age 50. About 6 in 10 cases of prostate cancer are found in men older than 65.

Race/ethnicity: Prostate cancer occurs more often in African-American men and in Caribbean men of African root than in men of other races. African-American men are also more than twice as likely to die of prostate cancer as white men. Prostate cancer occurs less often in Asian-American and Hispanic/Latino men than in non-Hispanic whites. The reasons for these racial and ethnic differences are not clear.

Geography: Prostate cancer is most common in North America, northwestern Europe, Australia, and on Caribbean islands. It is less common in Asia, Africa, Central America, and South America. The reasons for this are not clear. More intensive screening in some developed countries probably accounts for at least part of this difference, but other factors such as lifestyle differences are likely to be important as well.

Family history

Prostate cancer seems to run in some families, which suggests that in some cases there may be an inherited or genetic factor. (Still, most prostate cancers occur in men without a family history of it). Having a father or brother with prostate cancer more than doubles a man's risk of developing this disease. (The risk is higher for men who have a brother with the disease than for those who have a father with it). The risk is much higher for men with several affected relatives, particularly if their relatives were young when the cancer was found.

Gene changes:

Several inherited gene changes seem to raise prostate cancer risk, but they probably account for only a small percentage of cases overall.

Factors with less clear effect on prostate cancer risk:

Diet:

The exact role of diet in prostate cancer is not clear, but several factors have been studied. Men who eat a lot of red meat or high-fat dairy products appear to have a slightly higher chance of getting prostate cancer. These men also tend to eat fewer fruits and vegetables. Doctors aren't sure which of these factors is responsible for raising the risk. Some studies have suggested that men who consume a lot of calcium (through food or supplements) may have a higher risk of developing prostate cancer. Dairy foods (which are often high in calcium) might also increase risk. But most studies have not found such a link with the levels of calcium found in the average diet, and it's important to note that calcium is known to have other important health benefits.

Obesity:

Being obese (very overweight) does not seem to increase the overall risk of getting prostate cancer. Some studies have found that obese men have a lower risk of getting a low-grade (less dangerous) form of the disease, but a higher risk of getting more aggressive prostate cancer. Some studies have also found that obese men may be at greater risk for having more advanced prostate cancer and of dying from prostate cancer, but not all studies have found this.

Smoking:

Most studies have not found a link between smoking and getting prostate cancer. Some research has linked smoking to a possible small increased the risk of dying from prostate cancer, but this finding needs to be confirmed by other studies.

Chemical exposures:

There is some evidence that firefighters can be exposed to chemicals that may increase their risk of prostate cancer. A few studies have suggested a possible link between exposure to Agent Orange, a chemical used widely during the Vietnam War, and the risk of prostate cancer, although not all studies have found such a link. The Institute of Medicine considers there to be "limited/suggestive evidence" of a link between Agent Orange exposure and prostate cancer.

Inflammation of the prostate:

Some studies have suggested that prostatitis (inflammation of the prostate gland) may be linked to an increased risk of prostate cancer, but other studies have not found such a link. Inflammation is often seen in samples of prostate tissue that also contain cancer. The link between the two is not yet clear.

Vasectomy:

Some studies have suggested that men who have had a vasectomy (minor surgery to make men infertile) have a slightly increased risk for prostate cancer, but other studies have not found this. Research on this possible link is still under way, (American Cancer Society 2016[e]).

(2.6) The Signs and Symptoms of Prostate Cancer:

A patient with early prostate cancer may have the following signs and symptoms:

- A frequent or excessive need to urinate, during the day and/or at night
- Difficulty in starting, maintaining, or stopping the urine stream
- A weak or interrupted urine stream
- Straining to urinate
- Inability to urinate (urinary retention)
- Loss of control of urination that may be associated with coughing or laughing, a sudden urge to urinate, or without any forewarning
- Difficulty urinating when standing, requiring sitting during urination
- Pain with urination or ejaculation
- Blood in the urine or in the semen
- Abnormal rectal examination

Many symptoms of early cancer of the prostate can also be attributed to benign (noncancerous) conditions of the prostate including benign prostatic hypertrophy (BPH), or infection in the prostate gland or urinary system. Signs and symptoms of advanced prostate cancer that has already spread from the prostate gland to elsewhere in the body (called metastatic prostate cancer) include:

- Feeling of pain in the bones, especially the low back
- Unexplained weight loss
- Fatigue
- Increasing shortness of breath while doing activities previously well tolerated
- Low-impact fracture of bone(s) without a lot of trauma (or broken bone[s] from minor trauma), (American Cancer Society 2016[f]).

(2.7) Early Detection of Prostate Cancer:

Cancer early often allows for more treatment options. Some early cancers may have signs and symptoms that can be noticed, but that is not always the case. A doctor will usually do a blood test and/or physical examination to check the health of the prostate.

Blood Test (Prostate Specific Antigen (PSA) test): The result shows whether there is an increase in this specific protein. Depending on the result, you might need further investigation by a specialist. A high PSA test result does not necessarily mean cancer. Prostate diseases other than cancer can also cause a higher than normal PSA level.

Digital Rectal Examination (DRE): Because of where the prostate is located, the doctor inserts a gloved, lubricated finger into the rectum to check the size of the prostate and assess if there are any abnormalities. A normal DRE result does not rule out prostate cancer, (American Cancer Society 2016[g]).

Screening is testing to find cancer in people before they have symptoms. For some types of cancer, screening can help find cancers at an early stage, when they are likely to be easier to treat. Prostate cancer can often be found before symptoms start by testing the amount of prostate-specific antigen (PSA) in a man's blood. Another way to find prostate cancer is the digital rectal exam (DRE). If the results of either one of these tests are abnormal, further testing is often done to see if a man has cancer. If prostate cancer is found as a result of screening with the PSA test or DRE, it will probably be at an earlier, more treatable stage than if no screening were done. If the results of early detection tests – the (PSA) blood test and/or (DRE) – suggest that you might have prostate cancer, the doctor will do other tests, such as a transrectal ultrasound and a prostate biopsy to find out. There is no question that screening can help find many prostate cancers early. There are clearly both pros and cons to the prostate cancer screening tests in use today. So men should thinking about getting screened for prostate cancer should make informed decisions based on available information, discussion with their doctor, and their own views on the possible benefits, risks, and limits of prostate cancer screening.

Transrectal Ultrasound (TRUS): For this test, a small probe about the width of a finger is lubricated and placed in the rectum. The probe gives off sound waves that enter the prostate and create echoes. The probe picks up the echoes, and a computer turns them into a black and white image of the prostate. This procedure often takes less than 10 minutes and is done in a doctor's office or outpatient clinic. You will feel some pressure when the TRUS probe is placed in your rectum, but it is usually not painful. TRUS is useful in other situations as well. It can be used to measure the size of the prostate gland, which can help determine the PSA density and may also affect which treatment options a man has. TRUS is also used as a guide during some forms of prostate cancer treatment.

Prostate Biopsy: A biopsy is a procedure in which small samples of the prostate are removed and then looked at under a microscope. A core needle biopsy is the main method used to diagnose prostate cancer. It is usually done by an urologist. Using TRUS to "see" the prostate gland, the doctor quickly inserts a thin, hollow needle through the wall of the rectum and into the prostate. When the needle is pulled out, it removes a small cylinder (core) of prostate tissue. This is repeated several times. Most urologists will take about 12 core samples from different parts of the prostate, (American Cancer Society 2016[h]).The results from a prostate biopsy are usually given in the form of the Gleason score. On the simplest level, this scoring system assigns a number from 2 to 10 to describe how abnormal the cells appear under a microscope. A score of 2 to 4 means the cells still look very much like normal cells and pose little danger of spreading quickly. A score of 8 to 10 indicates that the cells have very few features of a normal cell and are likely to be aggressive. A score of 5 to 7 indicates intermediate risk, (Jaret, Peter 2010).

(2.8) Prostate Cancer Stages:

The stage (extent) of a prostate cancer is one of the most important factors in choosing treatment options and predicting a man's outlook for survival. The stage is based on:

- The prostate biopsy results (including the Gleason score)
- The blood PSA level at the time of diagnosis
- The results of any other exams or tests that were done to find out how far the cancer has spread.

The AJCC TNM staging system: A staging system is a standard way for the cancer care team to describe how far a cancer has spread. The most widely used staging system for prostate cancer is the American Joint Committee on Cancer (AJCC) TNM system. The TNM system for prostate cancer is based on 5 key pieces of information:

- The extent of the main (primary) tumor (T category)
- Whether the cancer has spread to nearby lymph nodes (N category)
- Whether the cancer has spread (metastasized) to other parts of the body (M category)
- The PSA level at the time of diagnosis
- The Gleason score, based on the prostate biopsy (or surgery)

There are 2 types of staging for prostate cancer:-

- The **clinical stage** is the doctor's best estimate of the extent of the disease, based on the results of the physical exam (including DRE), lab tests, prostate biopsy, and any imaging tests you have had.
- If you have surgery, the doctors can also determine the **pathologic stage**, which is based on results above, plus the results of the surgery. This means that if a patient has surgery, the stage of the cancer might actually change afterward (if cancer was found in a place it wasn't suspected). Pathologic staging is likely to be more accurate than clinical staging, as it gives the doctor a firsthand impression of the extent of the disease. Both types of staging use the same categories:

T Categories:

There are 4 categories for describing the local extent of a prostate tumor, ranging from T1 to T4:

T1: Your doctor can't feel the tumor or see it with imaging such transrectal ultrasound.

T2: Your doctor can feel the cancer with a digital rectal exam (DRE) or see it with imaging such as transrectal ultrasound, but it still appears to be confined to the prostate.

T3: The cancer has grown outside your prostate and may have grown into the seminal vesicles.

T4: The cancer has grown into tissues next to the prostate (other than the seminal vesicles), such as the urethral sphincter (a muscle that helps control urination), the rectum, the bladder, and/or the wall of the pelvis.

N Categories:

N categories describe whether the cancer has spread to nearby (regional) lymph nodes.

NX: Nearby lymph nodes were not assessed.

N0: The cancer has not spread to any nearby lymph nodes.

N1: The cancer has spread to one or more nearby lymph nodes.

M Categories:

M categories describe whether the cancer has spread to distant parts of the body. The most common sites of prostate cancer spread are to the bones and to distant lymph nodes, although it can also spread to other organs, such as the lungs and liver.

M0: The cancer has not spread beyond nearby lymph nodes.

M1: The cancer has spread beyond nearby lymph nodes, (American Cancer Society 2016[i]).

(2.9) Treating Prostate Cancer:

Depending on each case, treatment options for men with prostate cancer might include:

- Watchful waiting or active surveillance

- Surgery
- Radiation therapy
- Cryotherapy (cryosurgery)
- Hormone therapy
- Chemotherapy
- Vaccine treatment
- Bone-directed treatment

It's important to discuss all of the treatment options, including their goals and possible side effects, with the doctors to help make the decision that best fits patient's needs. Some important things to consider include:

- The stage and grade of the cancer
- Patient's age and expected life span
- Any other serious health conditions that patient has
- Patient feelings about the need to treat the cancer right away
- The likelihood that treatment will cure the cancer
- Patient feelings about the possible side effects from each treatment

The patient may feel that he must make a decision quickly, but it's important to give him time to understand the information he has just learned. It's also very important to ask questions if there is anything he is not sure about. Once the prostate cancer has been diagnosed and staged, the patient has a lot to think about before the choice of a treatment plan. It's important that he think carefully about each of his choices. The patient will want to weigh the benefits of each treatment option against the possible risks and side effects.

For most men with prostate cancer, treatment can remove or destroy the cancer. Completing treatment can be both stressful and exciting. For other men, the cancer may come back in other parts of the body or may never go away completely. These men may get hormone treatment or other therapies to help keep the cancer in check for as long as possible, (American Cancer Society 2016[j]).

Chapter 3

The Statistical Method

Chapter Three

(3.1) preface:

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables. Sometimes the dependent variable becomes discrete, taking two or more values. Logistic regression interested in such cases. The aim of using this method is to find the best fitting model to describe the relationship between an outcome or response variable and asset of predictor or explanatory variables called covariates. Other modeling approaches are possible also, but when the illness measure is dichotomous, the logistic regression is most popular modeling procedure use to analyze epidemiologic data, because it is easily used function and gives a clinically meaningful interpretation, (Hosmer, David W., Lemeshow, Stanley 2002).

(3.2) Simple Logistic Regression:

Logistic regression is a mathematical modeling approach used to describe a relationship of several explanatory (independent or predictor) variables to a dichotomous outcome (dependent or response) variable. In any regression problem the response variable is the mean value of the outcome variable given the value of predictor variables, as:

$$E(Y|X) = \beta_0 + \beta_1 X \rightarrow (3.1)$$

Where: β's are regression coefficients. This implies that the right side takes any value as x ranges between (-∞) to (+∞), but the left side is dichotomous data.

$0 \leq \pi(X) \leq 1$, where: $\pi(X) = E(Y|X)$.

The odds and natural logarithm solved this problem.

$$odds = \frac{p(y=1)}{p(y=0)} = \frac{\pi(x)}{1-\pi(x)}, 0 \leq \pi(x) \leq \infty \rightarrow (3.2)$$

By taking natural logarithm for both sides of the equation (3.2) will be as:

$$\log(odds) = log\frac{\pi(x)}{1-\pi(x)}, \qquad -\infty \leq log\frac{\pi(x)}{1-\pi(x)} \leq \infty \rightarrow (3.3)$$

Equation (3.3) called log transformation. The logit function is linear transformation in x:

$$logit\ (y) = \beta_0 + \beta_1 X \rightarrow (3.4)$$

Also can convert logit(y) back to the odds by exponentiation equation (3.4) as:

$$odds\ (y = 1) = EXP\ (\beta_0 + \beta_1 X) \rightarrow (3.5)$$

Similarly can convert odds back to probability that (y=1) by the formula:

$$p(y = 1) = \frac{odds(y = 1)}{[1 + odds(y = 1)]} \rightarrow (3.6)$$

The equation (3.6) will be as:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \rightarrow (3.7)$$

Or:

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \rightarrow (3.8)$$

This model was designed to give a probability of risk of individual to get an event or (disease). The curve of the logistic distribution said to be S-shaped, (O'Connell, Ann A. 2006).

(3.3) Assumptions of Logistic Regression:

1- Logistic regression doesn't assume a linear relationship between the outcome variable and explanatory variables, as linear regression does.
2- The dependent variable (DV) must be a dichotomy or binary, but in linear regression the DV must be continuous variable.

3- The independent variable (IV) need not be interval, nor normality distributed, nor linearity related, nor of equal variance within each group.
4- The categories or groups must be mutually exclusive and exhaustive that means a case only be in one group and every subject must be a member of one of the groups.
5- Larger samples are needed than for linear regression, because maximum likelihood coefficients are large sample estimates.
6- In linear regression an observation of the outcome variable expressed as :

$$Y = E(y|x) + e \rightarrow (3.9)$$

The error term (e) is an observation's deviation from conditional mean. The most common assumption is that e follows normal distribution with zero mean and some constant variance, which means the conditional distribution, will be normal distribution with mean $E(y|x)$ and constant variance. In a case that the outcome variable (DV) is dichotomous the value of outcome variable given x expressed as:

$$y = \pi(X) + e \rightarrow (3.10)$$

(e) has two possible values:

If y =1, then e =$1 - \pi(x)$, with probability$\pi(x)$.

If y =0, then e =$-\pi(x)$, with probability$1 - \pi(x)$.

So the binomial distribution describes the distribution of the error term with mean zero and variance equal$\pi(x)[1 - \pi(x)]$.

(3.4) Fitting the Logistic Regression Model:

As in linear regression we are trying to find best fitting line that represents data precisely. Because the DV (Y) in logistic regression can only range between 0 and 1, we cannot use the least square approach. The maximum likelihood (ML) used instead to find the function that will

maximize our ability to predict the probability of y based on given values of predictor variable (x), (Menard, Scott. 2002). This method provides the foundation for our approach to estimate unknown parameters which maximize the probability of obtaining the observed asset of data. If y coded as 0 or 1 and $p(y = 1|x) = \pi(x)$ and $(y = 0|x) = 1 - \pi(x)$, then the likelihood function written as:

$$l(\beta) = \prod_{i=0}^{n} \pi(x_i)^{y_i} \, [1 - \pi(x_i \,)]^{1-y_i} \rightarrow (3.11)$$

Taking natural logarithm:

$$\ln(l(\beta)) = \sum_{i=1}^{n} \left[y_i \ln \pi(x_i) + (1 - y_i) \ln\left(1 - \pi(x_i)\right) \right] \rightarrow (3.12)$$

We can find the ML estimate of β by differentiate equation (3.12) with respect to β_0 and β_1, then set the new equation equal to zero.

(3.5) Testing for the Significance of the Coefficients:

After estimating the coefficients, we concerned on assessment of the significance of the variable of statistical hypothesis to know whether the IV's in the model significantly related to the outcome variable (DV). One approach to testing for the significance of the coefficient of the variable in any model by comparing observed and predicted value of the response variable (outcome) by each of two models, model with predictor variable and model without.

In logistic regression the comparison of observed and predicted values by using the likelihood function is based on:

$$D = -2ln \left[\frac{liklihood\ of\ the\ fitted\ model}{liklihood\ of\ the\ saturated\ model} \right] \rightarrow (3.13)$$

Saturated model is a model with all coefficients under the study, and fitted model is a model with constant only. The expression between two brackets called likelihood ratio, this ratio is multiplying by (-2 ln) for obtaining form

that has known distribution. Chi- square test is used to assess significance of the ratio.

$$D = -2\sum_{i=1}^{n} y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1 - y_i) ln\left(\frac{1 - \hat{\pi}_i}{1 - y_i}\right) \rightarrow (3.14)$$

Where: $\hat{\pi}_i$=$\hat{\pi}_i(X)$.

The statistics D is called deviance, the change in D due to insertion of the predictor variable (ID) to the model expressed as:

G = D (model without of the variable) – D (model with the variable), (O'Connell, Ann A. 2006).

In logistic regression two hypothesis one of interest, the null hypothesis (H_0) states that all coefficients of regression take zero value, and the alternative hypothesis (H_1) states that at least one parameter not equal to zero, (Menard, Scott. 2002). There are two other tests used to determine the significance of the coefficients, these tests are; Wald test and Score test. The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter ($\hat{B}_1$), to an estimate of its standard error:

$$W = \frac{\hat{B}_1}{SE(\hat{B}_1)} \rightarrow (3.15)$$

This ratio under the hypothesis that ($B_1 = 0$) will follow standard normal distribution. After for significance a variable which doesn't require the maximum likelihood estimate of B_1 is the Score test. It is easy to use, but it cannot be obtained by software packages. The test statistics for the Score test (ST) is:

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1 - \bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}} \rightarrow (3.16)$$

(3.6) Confidence Interval Estimation:

Confidence interval (CI) is another way of estimation the parameters, slope or intercept based on their Wald tests. The endpoints of a $100(1-\alpha)$ % confidence interval for slope coefficients are:

$$\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \widehat{SE}\left(\hat{\beta}_1\right) \rightarrow (3.17)$$

And CI for the intercept as:

$$\hat{\beta}_0 \pm Z_{1-\frac{\alpha}{2}} \widehat{SE}\,\hat{\beta}_0 \rightarrow (3.18)$$

$\widehat{SE}$ (.): estimated standard error.

The estimator of the logit as:

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \rightarrow (3.19)$$

Then :

$$var\hat{g}(x) = var\left(\hat{\beta}_0\right) + var\left(\hat{\beta}_1\right) + 2x\widehat{cov}(\beta_0, \beta_1) \rightarrow (3.20)$$

So, the endpoints of a 100(1-α)% CI for logit are:

$$\hat{g}(x) \pm z_{1-\frac{\alpha}{2}} \widehat{SE}\ \hat{g}(x) \rightarrow (3.21)$$

(3.7) Multiple Logistic Regressions:

In this section we will introduce the logistic regression model with more than one IV. Assume there are (p) independent variables denoted by vector $\boldsymbol{x}$ as:

$\grave{\boldsymbol{x}} = (x_1, x_2, x_3, \dots, x_p)$, and $p(y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x})$, then the logit of the multiple logistic regression model and logistic regression model are given by the forms, respectively:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \rightarrow (3.22)$$

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} \rightarrow (3.23)$$

All the IVs must be interval scale, if one of IVs is discrete or nominal scale the method is to use a collection of design variables or dummy variables. Suppose that the j[th] IV is discrete or nominal scale and has k_j levels, then $k_j - 1$ design variables will be needed, denoted as D_{ji} thus the logit model written as:

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{i=1}^{k_j-1} \beta_{ji} D_{ji} + \cdots + \beta_p x_p \rightarrow (3.24)$$

(3.8) Fitting the Multiple Logistic Regression Model:

Assume we have a sample of n independent observations, the method used to estimate the vector of coefficients $\grave{\boldsymbol{\beta}} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$ is maximum likelihood. After differentiating the log of likelihood function with respect to the (p+1) coefficients, then (p+1) likelihood equations will be obtained as:

$$\frac{\partial^2 l(\beta)}{\partial {\beta_j}^2} = -\sum_{i=1}^{n} {x_{ij}}^2 \pi_i (1 - \pi_i) \rightarrow (3.25)$$

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_i} = -\sum_{i=1}^{n} x_{ij} x_{ji} \pi_i (1 - \pi_i) \rightarrow (3.26)$$

I ($\boldsymbol{\beta}$) is a (p+1)*(p+1) matrix contains the negative of the terms given in equations (3.25) and (3.26), called observed information matrix. The variance and covariance of the estimated coefficients are obtained from the inverse of this matrix [$\boldsymbol{I^{-1}(\beta)}$]**.** The information matrix $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}) = \grave{\boldsymbol{X}}\boldsymbol{V}\boldsymbol{X}$**,** where $\boldsymbol{X}$ is an n*(p+1) matrix and **V** is (n*n) diagonal matrix as:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 \ \dots & 0 \\ 0 & \ddots & \vdots \\ 0 & 0 \ \dots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

(3.9) Testing for the Significance of the Model:

The likelihood ratio test for overall significance of the p coefficients for the variables in the model is performed. This test based on the G statistics:

$$G = -2ln\left[\frac{liklihood\ without\ the\ variables}{liklihood\ with\ the\ variables}\right]$$

Under the null hypothesis that the p coefficients for the covariates are zeros, and the alternative hypothesis states that there is at least one coefficient is different from zero, or all coefficients are non zeros. The distribution of G statistics is chi-square with p degree of freedom. In this situation we have a vector $\widehat{\boldsymbol{\beta}}$ contains (p+1) parameters. Wald test used to test the significance of coefficients, under the null hypothesis that all the (p+1) coefficients are zeros.

$$\boldsymbol{W} = \hat{\hat{\boldsymbol{\beta}}}[\widehat{var}(\widehat{\boldsymbol{\beta}})]^{-\mathbf{1}}\widehat{\boldsymbol{\beta}} \sim \mathcal{X}^2(p+1)$$

$$\boldsymbol{W} = \hat{\hat{\boldsymbol{\beta}}}\big(\grave{\boldsymbol{X}}\boldsymbol{V}\boldsymbol{X}\big)\widehat{\boldsymbol{\beta}} \rightarrow (3.27)$$

(3.10) Confidence Interval Estimation:

The general expression for the estimator of the logit for a model containing p covariates is:

$$\hat{g}(\boldsymbol{X}) = \grave{\boldsymbol{X}}\widehat{\boldsymbol{\beta}} \rightarrow \grave{(}3.28)$$

Where: $\hat{\hat{\boldsymbol{\beta}}} = (\beta_0, \beta_1, \dots, \beta_p)$ and $\hat{\grave{\boldsymbol{X}}} = (x_0, x_1, x_2, \dots, x_p)$, represents a set of values of the p covariates in the model, where$x_0 = 1$. The estimator expression of variance of the coefficients, from information matrix is obtained by:

$$\widehat{var}\big(\widehat{\boldsymbol{\beta}}\big) = (\acute{\boldsymbol{X}}\boldsymbol{V}\boldsymbol{X})^{-\mathbf{1}} \rightarrow (3.29)$$

Then:

$$\widehat{var}\left(\widehat{\boldsymbol{g}}(\boldsymbol{X})\right) = \left[\acute{\boldsymbol{X}}\widehat{\boldsymbol{var}}(\widehat{\boldsymbol{\beta}})\boldsymbol{X}\right] \rightarrow (3.30)$$

$$\widehat{var}(\widehat{\boldsymbol{g}}(\boldsymbol{X})) = [\acute{\boldsymbol{X}}(\acute{\boldsymbol{X}}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}] \rightarrow (3.31)$$

Fortunately, all good logistic regression software packages provide the option for the user to create a new variable containing the estimated values for equation (3.31) or the standard error for all subjects in the data set. This allows the user to calculate fitted values and confidence interval estimates. We discussed confidence interval for the coefficients and logit for the simple logistic regression model in section (3.6). The methods used for confidence interval estimators for multiple variable models are essentially the same.

(3.11) Interpretation of the Fitted Logistic Regression Model:

In the previous sections we discussed the methods for fitting and testing for the significance of the logistic regression model. After fitting a model the emphasis shifts from the computation and assessment of the significance of the estimated coefficients to the interpretation of their values. The interpretation of any fitted model requires that we be able to draw practical inferences from estimated coefficients in the model. The estimated coefficients for the IVs represent the slope of a function of the DV per unit of change in the IV. Thus the interpretation involves two issues: determining the functional relationship between the DV and the IV, and defining the unit of change for the IV. The first step is to determine the link function. It is a function of the DV yields a linear function of the IVs. In the logistic regression the link function is the logit transformation as:

$$g(x) = \ln\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x \rightarrow (3.32)$$

For linear regression model, the slope coefficient β_1 is equal to the difference between the value of the DV at (x+1) and the value of the DV at x, so:

If $y(x) = \beta_0 + \beta_1 x$, and $y(x+1) = \beta_0 + \beta_1(x+1)$, then:

$$\beta_1 = y(x+1) - y(x) \rightarrow (3.33)$$

In this case the interpretation of the coefficient is the resulting change in the measurement scale of the DV for a unit change in the IV. In the logistic regression, the slope coefficient represents the change in the logit corresponding to a change of one unit in the IV. In the following section we consider the interpretation of the coefficients for a univariate logistic regression model for each of the possible measurement scales of the IV.

(3.12) Dichotomous independent variable:

We consider that the IV is nominal scale and dichotomous, and we assume that the IV x is coded as either zero or one. The difference in the logit for a subject with x=1 and x=0 is:

$$g(1) - g(0) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1 \rightarrow (3.34)$$

To interpret this result we need to discuss a measure of association called odds ratio. The odds of the outcome being present among individuals with x=1 is defined as:

$$odds(y = 1) = \frac{\pi(1)}{1 - \pi(1)} \rightarrow (3.35)$$

Similarly, the odds of the outcome being present among individuals with x=0 is:

$$odds(y = 0) = \frac{\pi(0)}{1 - \pi(0)} \rightarrow (3.36)$$

The odds ratio (OR) is defined as the ratio of the odds for x=1 to the odds for x=0 as:

$$OR = \frac{\pi(1)/[1 - \pi(1)]}{\pi(0)/[1 - \pi(0)]} = e^{\beta_1} \rightarrow (3.37)$$

The log of odds ratio can provide the estimated coefficients. Equation (3.37) represents the relationship between the odds ratio and the regression coefficient; it is the foundation reason why logistic regression has proven to be such a powerful analytic research tool. The odds ratio, OR, is usually the parameter of interest in a logistic regression due to its ease of interpretation. However, its estimate, OR tends to have a distribution that is skewed. The skewness of the sampling distribution of OR is due to the fact that possible values range between 0 and∞, with the null value equaling 1. In theory, for large enough sample sizes, the distribution of OR is normal. Hence, inferences are usually based on the sampling distribution of

$$\ln(\widehat{\mathrm{OR}}) = \hat{\beta}_1 \rightarrow (3.38)$$

It tends to follow a normal distribution for much smaller sample sizes. A 100(1-α)% confidence interval (CI) estimate for the odds ratio is obtained by first calculating the endpoints of a confidence interval for the coefficient, $\hat{\beta}_1$, and then exponentiating these values. In general, the endpoints are given by the expression:

$$exp\left[\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} * \widehat{SE}(\hat{\beta}_1)\right] \rightarrow (3.39)$$

The interpretation given for the odds ratio is based on the fact that in many instances it approximates a quantity called the relative risk (RR). This

parameter is equal to the ratio $\pi(1)/\pi(0)$, the odds ratio approximates the relative risk if:$[1 - \pi(0)]/[1 - \pi(1)] \approx 1$.

The estimate of odds ratio discussed in equation (3.38) is accurate when the independent variable is coded as 0 and 1. Other coding may require calculating the value of the logit difference for the specific coding, exponentiation of the difference obtains the odds ratio. The estimate of the log of the odds ratio for any independent variable at two different levels, say x=a, versus x=b, is the difference between the estimated logit computed at these two values:

$$\begin{aligned}\ln(\widehat{\mathrm{OR}}) &= \hat{g}(x=a) - \hat{g}(x=b)\\ &= \hat{\beta}_0 + a\,\hat{\beta}_1 - (\hat{\beta}_0 + b\,\hat{\beta}_1)\\ &= \hat{\beta}_1(a-b)\end{aligned}$$

The estimate of odds ratio is obtained by exponentiating the logit difference:

$$\widehat{OR}(a,b) = \exp\,[\hat{\beta}_1(a-b)]$$

The notation $\widehat{OR}(a,b)$ is used to represent the odds ratio,

$$\widehat{OR} = \frac{\hat{\pi}(x=a)/[1-\pi(x=a)]}{\hat{\pi}(x=b)/[1-\pi(x=b)]} \rightarrow (3.40)$$

When a=1 and b=0 the $\widehat{OR} = \widehat{OR}(1,0)$.

The "zero-one" coding used so far in this section is frequently referred to as reference cell coding. The reference cell method typically assigns the value of zero to the lower code for *x* and one to the higher code. Another coding method is frequently referred to as deviation from means coding. This method assigns the value of -1 to the lower code, and a value of 1 to the higher code. This method used to estimate the odds ratio of x=a and x=b,

$$\begin{aligned}ln\left[\widehat{OR}(a,b)\right] &= \hat{g}(a) - \hat{g}(b)\\ &= \hat{g}(D=1) - \hat{g}(D=-1)\\ &= [\hat{\beta}_0 + \hat{\beta}_1 * (D=1)] - [\hat{\beta}_0 + \hat{\beta}_1 * (D=-1)]\\ &= 2\hat{\beta}_1\end{aligned}$$

$$\therefore \widehat{OR} = exp(2\hat{\beta}_1).$$

The method of coding also influences the calculation of the endpoints of the confidence interval, the estimated standard error needed for confidence interval estimation is $\widehat{SE}(2\hat{\beta}_1)$ which is $2\widehat{SE}(\hat{\beta}_1)$. Then the endpoints of the CI are

$$exp\left[2\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} * 2\widehat{SE}(\hat{\beta}_1)\right] \rightarrow (3.41)$$

In general:

$$exp\left[\hat{\beta}_1(a-b) \pm Z_{1-\frac{\alpha}{2}} * |a-b|\widehat{SE}(\hat{\beta}_1)\right]$$

where $|a-b|$ is the absolute value of (a-b).

(3.13) Polychotomous Independent Variable:

Suppose that instead of two categories the independent variable has $k > 2$ distinct values. a set of design variables should be taken to represent the categories of the variable. This section presents methods for creating design variables for polychotomous independent variables. The method for specifying the design variables involves setting all of them equal to zero for the reference group, and then setting a single design variable equal to 1 for each of the other groups. this method is usually referred to as reference cell coding and is the default method in many packages. A comment about the estimated standard errors may be helpful at this point. In the univariate case the estimates of the standard errors found in the logistic regression output are identical to the estimates obtained using the cell frequencies from the contingency table.

$$\widehat{SE}(\beta_1) = \left[\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right]^{0.5} \rightarrow (3.42)$$

Confidence limits for odds ratios are obtained by using the same approach used for a dichotomous variable. First step is to compute the confidence limits for the log odds ratio (the logistic regression coefficient) and then exponentiate these limits to obtain limits for the odds ratio. In general, the limits for a $100(1-\alpha)\%$ CIE for the coefficient are of the form:

$$\widehat{\beta_J} \pm Z_{1-\frac{\alpha}{2}} * \widehat{SE}(\widehat{\beta_J}) \rightarrow (3.43)$$

The corresponding limits for the odds ratio, obtained by exponentiating these limits, are as follows:

$$exp\left[\widehat{\beta_J} \pm Z_{1-\frac{\alpha}{2}} * \widehat{SE}(\widehat{\beta_J})\right] \rightarrow (3.44)$$

Reference cell coding is the most commonly employed coding method appearing in the literature. The primary reason for the widespread use of this method is the interest in estimating the risk of an "exposed" group relative to that of a "control" or "unexposed" group. Second method of coding design variables is called deviation from means coding. This coding expresses effect as the deviation of the "group mean" from the "overall mean." In the case of logistic regression, the "group mean" is the logit for the group and the "overall mean" is the average logit over all groups. This method of

coding is obtained by setting the value of all the design variables equal to -1 for one of the categories, and then using the 0, 1 coding for the remainder of the categories. The interpretation of the estimated coefficients is not as easy or clear as in the situation when a reference group is used. Exponentiation of the estimated coefficients yields the ratio of the odds for the particular group to the geometric mean of the odds. The estimated coefficients obtained using deviation from means coding may be used to estimate the odds ratio for one category relative to a reference category. The equation for the estimate is more complicated than the one obtained using the reference cell coding. It should be apparent that, if the objective is to obtain odds ratios, use of deviation from means coding for design variables is computationally much more complex than reference cell coding.

In summary, we have shown that discrete nominal scale variables are included properly into the analysis only when they have been recoded into design variables. The particular choice of design variables depends on the application, though the reference cell coding is the easiest to interpret.

(3.14) Continuous independent variable:

When a logistic regression model contains a continuous independent variable, interpretation of the estimated coefficient depends on how it is entered into the model and the particular units of the variable. The logit is linear in the variable was assumed for purposes of developing the method to interpret the coefficient for a continuous variable. Under this assumption, the equation for the logit is

$$g(x) = \beta_0 + \beta_1 x$$

It follows that the slope coefficient β_1 gives the change in the log odds for an increase of **1** unit in x, that is $\beta_1 = g(x+1) - g(x)$, for any value of x. The log odds ratio for a change of c units in x is obtained from the logit difference $g(x+c) - g(x) = c\beta_1$ and the associated odds ratio is obtained by exponentiating this logit difference,

$OR(c) = OR(x+c, x) = \exp(c\,\beta_1)$.

The estimate may be obtained by replacing β_1 with its maximum likelihood estimate $\widehat{\beta_1}$. An estimate of the standard error needed for confidence interval estimation is obtained by multiplying the estimated standard error of $\widehat{\beta_1}$ by c. Hence the endpoints of the $100(1-\alpha)\%$ CI estimate of $OR(c)$ are:

$$exp\left[c\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} * C\ \widehat{SE}(\hat{\beta}_1)\right] \rightarrow (3.45)$$

Since both the point estimate and endpoints of the confidence interval depend on the choice of c, the particular value of c should be clearly specified. In summary, the interpretation of the estimated coefficient for a continuous variable is similar to that of nominal scale variables: an estimated log odds ratio. The primary difference is that a meaningful change must be defined for the continuous variable.

(3.15) Model Building Strategies and Methods for Logistic Regression:

The previous sections focused on estimating, testing, and interpreting the coefficients in a logistic regression model. In many situations there are many independent variables that could potentially be included in the model. Hence, developed a strategy and associated methods will be used for handling these more complex situations. The goal of any method is to select those variables that result in a "best" model within the scientific context of the problem. To achieve this goal a basic plan for selecting the variables for the model should be illustrated and a set of methods for assessing the adequacy of the model both in terms of its individual variables and its overall fit should be clarified, (Hosmer, David W., Lemeshow, Stanley 2002).

(3.16) Variable Selection Methods:

The criteria for including a variable in a model may vary from one problem to the next and from one scientific discipline to another. The traditional approach to statistical model building involves seeking the most parsimonious model that still explains the data. There are several steps one can follow to aid in the selection of variables for a logistic regression model. The process of model building is quite similar to the one used in linear regression, (Hosmer, David W., Lemeshow, Stanley 2002). Variable selection is intended to select the best subset of predictors. Some important points must to keep it in mind:

1- Explanation the data in the simplest way , redundant predictors should be removed. The principle of Occam's Razor, (Heylighen, F. 1997) states that among several plausible explanations for a phenomenon, the simplest is best. Applied to regression analysis, this implies that the smallest model that fits the data is best.
2- Unnecessary predictors will add noise to the estimation of other quantities that we are interested in.

3- Collinearity is caused by having too many variables trying to do the same job.
4- Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.
5-

(3.17) Stepwise Procedures:

Another approach to variable selection is to use a stepwise method in which variables are selected either for inclusion or exclusion from the model in a sequential fashion based solely on statistical criteria. There are two main versions of the stepwise procedure: forward selection and backward elimination. The stepwise approach is useful and in that it builds models in a sequential fashion and it allows for the examination of a collection of models which might not otherwise have been examined. Best subsets selection" is a selection method that has not been used extensively in logistic regression. With this procedure a number of models containing one, two, three variables, and so on, are examined to determine which are considered the "best" according to some specified criteria.

1- Backward Elimination:

This is the simplest of all variable selection procedures and can be easily implemented without special software. In situations where there is a complex hierarchy, backward elimination can be run manually while taking account of what variables are eligible for removal.

I. Start with all the predictors in the model
II. Remove the predictor with highest p-value greater than α.
III. Refit the model and go to II
IV. Stop when all p-values are less than α.

2- Forward Selection:

This just reverses the backward method.

I. Start with no variables in the model.
II. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α.
III. Continue until no new predictors can be added.

3- Stepwise Regression:

This is a combination of backward elimination and forward selection. This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later. At each stage a variable may be added or removed and there are several variations on exactly how this is done. Stepwise procedures are relatively cheap computationally but they do have some drawbacks.

I. Because of the .one-at-a-time, nature of adding/dropping variables, it's possible to miss the optimal model.
II. The removal of less significant predictors tends to increase the significant of the remaining predictors. This effect leads one to overstate the importance of the remaining predictors.
III. The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest. With any variable selection method, it is important to keep in mind that model selection cannot be divorced from the underlying purpose of the investigation. Variable selection tends to amplify the statistical significance of the variables that stay in the model. Variables that are dropped can still be correlated with the response. It would be wrong to say these variables are unrelated to the response; it's just that they provide no additional explanatory effect beyond those variables already included in the model.
IV. Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes.

4- Criterion-based procedures:

If there are p potential predictors, then there are 2^P possible models. We fit all these models and choose the best one according to some criterion. Clever algorithms such as the branch-and bound method can avoid actually fitting all the models, only likely candidates are evaluated. Some criteria are:

1- The Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) are some other commonly used criteria. In general,

$$AIC = -2loglikelihood + 2p \rightarrow (3.46)$$

$$BIC = -2loglikelihood + plogn \rightarrow (3.47)$$

2- Adjusted $\bar{R}^2$ recall that $R^2 = RSS/TSS$. Adding a variable to a model can only decrease the RSS and so only increase the $\bar{R}^2$ so $\bar{R}^2$ by itself is not a good criterion because it would always choose the largest possible model, (Weisberg, Sandford 2005).

(3.18) Assessing the Fit of the Model:

This section illustrates how effectively the model we have describes the outcome variable. This is referred to as its goodness-of-fit, (Hosmer, David W., Lemeshow, Stanley 2002). As in linear regression, goodness of fit in logistic regression attempts to get at how well a model fits the data. It is usually applied after a final model" has been selected, (Goodness of Fit in

Logistic Regression 2013). There are some specific ideas about what it means to say that a model fits, to assess the goodness-of-fit of the model. Suppose we denote the observed sample values of the outcome variable in vector form as y where $\grave{y} = (y_1,\ y_2,\ y_3, \ldots., y_n)$. Denote to the values predicted by the model, or fitted values, as $\hat{y}$ where $\hat{\hat{y}} = (\hat{y}_1,\ \ldots, \hat{y}_n)$. We conclude that the model fits if summary measures of the distance between y and $\hat{y}$ are small and the contribution of each pair $(y_i\ , \hat{y}_i)$, i= 1,2,…,n to the summary measures is unsystematic and is small relative to the error structure of the model. Thus, a complete assessment of the fitted model involves both the calculation of summary measures of the distance between y and y, and a thorough examination of the individual components of these measures. Much of the goodness of fit literature is based on hypothesis testing of the following type:

H_0: model is exactly correct

H_1: model is not exactly correct

When the model building stage has been completed, a series of logical steps may be used to assess the fit of the model. The components of the proposed approach are

(1) Computation and evaluation of overall measures of fit

(2) Examination of the individual components of the summary statistics, often graphically.

(3) Examination of other measures of the difference or distance between the components of y and $\hat{y}$.

(3.19) Summary Measures of Goodness of Fit:

- Chi-square goodness of fit tests and deviance
- Hosmer-Lemeshow tests
- Classification tables
- ROC curves
- Logistic regression R^2
- Model validation via an outside data set or by splitting a data set, (Hosmer, David W., Lemeshow, Stanley 2002).

(3.20) The Chi-Square Distribution and the Analysis of Frequencies:

The chi-square distribution is the most frequently employed statistical technique for the analysis of count or frequency data. The chi-square distribution may be derived from normal distributions. Suppose that from a normally distributed random variable *Y* with mean μ and variance σ^2 we randomly and independently select samples of size $n = 1$, each value selected may be transformed to the standard normal variable *z* by the familiar formula:

$$z_i = \frac{y_i - \mu}{\sigma}$$

Each value of *z* may be squared to obtain z^2. When we investigate the sampling distribution of z^2, we find that it follows a chi-square distribution with 1 degree of freedom.
That is,

$$\chi^2(1) = \left(\frac{y - \mu}{\sigma}\right)^2 = z^2 \rightarrow (3.48)$$

Now suppose that we randomly and independently select samples of size $n = 2$ from the normally distributed population of Y values. Within each sample we may transform each value of *y* to the standard normal variable z and square as before. If the resulting values of z^2 for each sample are added, we may designate this sum by

$$\chi^2(2) = \left(\frac{y_1 - \mu}{\sigma}\right)^2 + \left(\frac{y_2 - \mu}{\sigma}\right)^2 = {z_1}^2 + {z_2}^2 \rightarrow (3.49)$$

since it follows the chi-square distribution with 2 degrees of freedom, the number of independent squared terms that are added together. The procedure may be repeated for any sample size *n*. The sum of the resulting z^2 values in each case will be distributed as chi-square with *n* degrees of freedom. In general:

$$\chi^2(n) = {z_1}^2 + {z_2}^2 + \cdots + {z_n}^2 \rightarrow (3.50)$$

The above summation follows the chi-square distribution with *n* degrees of freedom. The mathematical form of the chi-square distribution is as follows:

$$f(u) = \frac{1}{\left(\frac{k}{2}-1\right)!\,(2)^{\frac{k}{2}}} * u^{(k/2)-1} * e^{-(u/2)}, u > 0 \rightarrow (3.51)$$

where *e* is the irrational number 2.71828 . . . and *k* is the number of degrees of freedom. The variable *u* is usually designated by the Greek letter chi (χ) and, hence, the distribution is called the chi-square distribution. The mean and variance of the chi-square distribution are *k* and 2*k*, respectively. The chi-square distributions is skewed to right because all values between 0 and ∞.It cannot take on negative values, since it is the sum of values that have

been squared. A final characteristic of the chi-square distribution worth noting is that the sum of two or more independent chi-square variables also follows a chi-square distribution.

(3.20.1)Types of Chi-Square Tests:

The chi-square distribution in this section concerned about testing hypotheses where the data available for analysis are in the form of frequencies. These hypothesis testing procedures are discussed under the topics of tests of goodness-of-fit, tests of independence, and tests of homogeneity. The chi-square statistic is most appropriate for use with categorical variables. The quantitative data used in the computation of the test statistic are the frequencies associated with each category of the one or more variables under study. There are two sets of frequencies with which we are concerned, observed frequencies and expected frequencies. The observed frequencies are the number of subjects or objects in our sample that fall into the various categories of the variable of interest. Expected frequencies are the number of subjects or objects in our sample that we would expect to observe if some null hypothesis about the variable is true.

(3.20.2) The Chi-Square Test Statistic

The test statistic for the chi-square tests we discuss in this section is:

$$X^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] \rightarrow (3.52)$$

When the null hypothesis is true, X^2 is distributed approximately as χ^2 with degrees of freedom $k - r$. In determining the degrees of freedom, k is equal to the number of groups for which observed and expected frequencies are available, and r is the number of restrictions or constraints imposed on the given comparison. A restriction is imposed when we force the sum of the expected frequencies to equal the sum of the observed frequencies, and an additional restriction is imposed for each parameter that is estimated from the sample. The O_i is the observed frequency for the i^{th} category of the variable of interest, and E_i is the expected frequency (given that H_0 is true) for the i^{th} category. The quantity X^2 is a measure of the extent to which, in a given situation, pairs of observed and expected frequencies agree. As we will see, the nature of X^2 is such that when there is close agreement between observed and expected frequencies it is small, and when the agreement is poor it is large. Consequently, only a sufficiently large value of X^2 will cause rejection of the null hypothesis. If there is perfect agreement between the observed frequencies and the frequencies that one would expect, given that $H_0$is true, the term $O_i - E_i$in Equation (3.52) will be equal to zero for

each pair of observed and expected frequencies. Such a result would yield a value of X^2 equal to zero, and we would be unable to reject H_0 . When there is disagreement between observed frequencies and the frequencies one would expect given that H_0 is true, at least one of the $O_i - E_i$terms in Equation (3.52) will be a nonzero number. In general, the poorer the agreement between O_i and E_i the greater or the more frequent will be these nonzero values. As noted previously, if the agreement between the O_i and the E_i is sufficiently poor (resulting in a sufficiently large X^2 value,) we will be able to rejectH_0.

When there is disagreement between a pair of observed and expected frequencies, the difference may be either positive or negative, depending on which of the two frequencies is the larger. Since the measure of agreement X^2 is a sum of component quantities whose magnitudes depend on the difference$O_i - E_i$, positive and negative differences must be given equal weight. This is achieved by squaring each $O_i - E_i$ difference. Dividing the squared differences by the appropriate expected frequency converts the quantity to a term that is measured in original units. Adding these individual $(O_i - E_i)^2/E_i$ terms yields X^2, a summary statistic that reflects the extent of the overall agreement between observed and expected frequencies.

The quantity $\sum[(O_i - E_i)^2/E_i]$ will be small if the observed and expected frequencies are close together and will be large if the differences are large. The computed value of X^2 is compared with the tabulated value $\chi^2$of with $k - r$ degrees of freedom. The decision rule, then, is: Reject H_0 if X^2 is greater than or equal to the tabulated for the chosen value of α .

(3.20.3) Tests of Goodness of Fit:

A goodness-of-fit test is appropriate when one wishes to decide if an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution. We may, for example, wish to determine whether or not a sample of observed values of some random variable is compatible with the hypothesis that it was drawn from a population of values that is normally distributed. The procedure for reaching a decision consists of placing the values into mutually exclusive categories or class intervals and noting the frequency of occurrence of values in each category. We then make use of our knowledge of normal distributions to determine the frequencies for each category that one could expect if the sample had come from a normal distribution. If the discrepancy is of such magnitude that it could have come about due to chance, we conclude that the sample may have come from a normal distribution. In a similar manner, tests of

goodness-of-fit may be carried out in cases where the hypothesized distribution is the binomial, the Poisson, or any other distribution.

(3.20.4) Tests of Independence:

Another, and perhaps the most frequent, use of the chi-square distribution is to test the null hypothesis that two criteria of classification, when applied to the same set of entities, are independent. We say that two criteria of classification are independent if the distribution of one criterion is the same no matter what the distribution of the other criterion.

(3.20.5) The Contingency Table:

The classification, according to two criteria, of a set of entities, say, people, can be shown by a table in which the *r* rows represent the various levels of one criterion of classification and the *c* columns represent the various levels of the second criterion. Such a table is generally called a contingency table. We will be interested in testing the null hypothesis that in the population the two criteria of classification are independent. If the hypothesis is rejected, we will conclude that the two criteria of classification are not independent. A sample of size n will be drawn from the population of entities, and the frequency of occurrence of entities in the sample corresponding to the cells formed by the intersections of the rows and columns along with the marginal totals will be displayed in a table such as Table (3.20.1).

In general, for calculating the expected frequencies for a given cell(i^{th} row and i^{th} column), we multiply the total of the row in which the cell is located by the total of the column in which the cell is located and divide the product by the grand total, as:

$$E_{ii} = \frac{(n_{i.})(n_{.i})}{n} \rightarrow (3.53)$$

The expected frequencies and observed frequencies are compared. If the discrepancy is sufficiently small, the null hypothesis is tenable. If the discrepancy is sufficiently large, the null hypothesis is rejected, and we conclude that the two criteria of classification are not independent. The X^2 shown in equation (3.52) is distributed approximately as with χ^2 with (r-1)(c-1) degrees of freedom when the null hypothesis is true. If the computed value of X^2 is equal to or larger than the tabulated value of χ^2 for some α, the null hypothesis is rejected at the level α of significance.

Table (3.1): Two-Way Classification of a Sample of Entities

Second criterion of classification level	First criterion of classification level					
	1	2	3	. . .	C	Total
1	n_{11}	n_{12}	n_{13}	. . .	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	. . .	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	. . .	n_{3c}	$n_{3.}$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
R	n_{r1}	n_{r2}	n_{r3}		n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	. . .	$n_{.n}$	n

Source: Wayne, W. Daniel 2009.

(3.20.6) The 2×2 Contingency Table:

Sometimes each of two criteria of classification may be broken down into only two categories, or levels. When data are cross classified in this manner, the result is a contingency table consisting of two rows and two columns. Such a table is commonly referred to as a 2×2 table. The value of X^2 may be computed by first calculating the expected cell frequencies as in equation (3.52). In the case of 2×2 a contingency table, however, may be calculated by the following shortcut formula:

$$X^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)} \rightarrow (3.54)$$

Where *a*, *b*, *c*, and *d* are the observed cell frequencies. When we apply the (r-1)*(c-1) rule for finding degrees of freedom to a 2×2 table, the result is 1 degree of freedom. The problems of how to handle small expected frequencies and small total sample sizes may arise in the analysis of 2×2 contingency tables. The χ^2 test should not be used if any of the expected frequency is less than 5.

Yates's Correction:

The observed frequencies in a contingency table are discrete and thereby give rise to a discrete statisticX^2, which is approximated by the χ^2

distribution, which is continuous. Yates, F. (1934) proposed a procedure for correcting for this in the case of 2×2 tables. The correction, as shown in Equation (3.55) consists of subtracting half the total number of observations from the absolute value of the quantity $(ad - bc)$ before squaring, as:

$$X^2{}_{corrected} = \frac{n(|ad - bc| - 0.5n)^2}{(a + c)(b + d)(a + b)(c + d)} \rightarrow (3.55)$$

It is generally agreed that no correction is necessary for larger contingency tables. Although Yates's correction for 2×2 tables has been used extensively in the past, more recent investigators have questioned its use. As a result, some practitioners recommend against its use.

Characteristics of Tests of Independence:

The characteristics of a chi-square test of independence that distinguish it from other chi-square tests are as follows:

1. A single sample is selected from a population of interest, and the subjects or objects are cross-classified on the basis of the two variables of interest.
2. The rationale for calculating expected cell frequencies is based on the probability law, which states that if two events (here the two criteria of classification) are independent, the probability of their joint occurrence is equal to the product of their individual probabilities.
3. The hypotheses and conclusions are stated in terms of the independence (or lack of independence) of two variables.

(3.20.7) Tests of Homogeneity:

In previous sections is that, in each case, the total sample was assumed to have been drawn before the entities were classified according to the two criteria of classification. That is, the observed number of entities falling into each cell was determined after the sample was drawn. As a result, the row and column totals are chance quantities not under the control of the investigator. We think of the sample drawn under these conditions as a single sample drawn from a single population. On occasion, however, either row or column totals may be under the control of the investigator; that is, the investigator may specify that independent samples be drawn from each of several populations. In this case, one set of marginal totals is said to be fixed, while the other set, corresponding to the criterion of classification applied to the samples, is random. The former procedure, as we have seen, leads to a chi-square test of independence. The latter situation leads to a chi square test of homogeneity. The two situations not only involve different sampling procedures; they lead to different questions and null hypotheses.

The test of independence is concerned with the question: Are the two criteria of classification independent? The homogeneity test is concerned with the question: Are the samples drawn from populations that are homogeneous with respect to some criterion of classification? In the latter case the null hypothesis states that the samples are drawn from the same population. Equation (3.53) can be used to calculate the expected frequencies to each cell and same chi-square statistic represented in equation (3.52) can be used to reject H_0 if X^2 is greater than or equal to the tabulated for the chosen value of α .

In summary, the chi-square test of homogeneity has the following characteristics:

1. Two or more populations are identified in advance, and an independent sample is drawn from each.
2. Sample subjects or objects are placed in appropriate categories of the variable of interest.
3. The calculation of expected cell frequencies is based on the rationale that if the populations are homogeneous as stated in the null hypothesis, the best estimate of the probability that a subject or object will fall into a particular category of the variable of interest can be obtained by pooling the sample data.
4. The hypotheses and conclusions are stated in terms of homogeneity (with respect to the variable of interest) of populations, the null hypothesis states; H_0: $P_1 = P_2$.

The chi-square test of homogeneity for the two-sample case provides an alternative method for testing the null hypothesis that two population proportions are equal. To test H_0: $P_1 = P_2$ against H_A: $P_1 \neq P_2$ by means of the statistic:

$$z = \frac{\left(\widehat{\mathrm{P}_1} - \widehat{\mathrm{P}_2}\right) - (\mathrm{P}_1 - \mathrm{P}_1)}{\sqrt{\left[\frac{\overline{\mathrm{P}}(1-\overline{\mathrm{P}})}{\mathrm{n}_1} + \frac{\overline{\mathrm{P}}(1-\overline{\mathrm{P}})}{\mathrm{n}_2}\right]}} \rightarrow (3.56)$$

Where $\bar{P}$ is obtained by pooling the data of the two independent samples available for analysis.

(3.21) The Fisher Exact Test:

Sometimes we have data that can be summarized in a 2×2 contingency table, but these data are derived from very small samples. The chi-square test is not an appropriate method of analysis if minimum expected frequency requirements are not met. If, for example, *n* is less than 20or if *n* is between

20 and 40 and one of the expected frequencies is less than 5, the chi-square test should be avoided. A test that may be used when the size requirements of the chi-square test are not met is Fisher exact test. It is called exact because, if desired, it permits us to calculate the exact probability of obtaining the observed results or results that are more extreme.

When we use the Fisher exact test, we arrange the data in the form of a 2×2 contingency table as in table (3.21.1) .We arrange the frequencies in such a way that $A > B$ and choose the characteristic of interest so that $a/A > b/B$. Some theorists believe that Fisher's exact test is appropriate only when both marginal totals of Table (3.21.1) are fixed by the experiment. This specific model does not appear to arise very frequently in practice. Many experimenters, therefore, use the test when both marginal totals are not fixed.

Table (3.2): A 2×2 Contingency Table For The Fisher Exact Test

Sample	With Characteristic	Without Characteristic	Total
1	A	A-a	A
2	B	B-b	B
Total	a +b	A+B-a-b	A+B

Source: Wayne, W. Daniel 2009.

Assumptions:

The following are the assumptions for the Fisher exact test.

1. The data consist of A sample observations from population 1 and B sample observations from population 2.
2. The samples are random and independent.
3. Each observation can be categorized as one of two mutually exclusive types.

Hypothesis:

The following are the null hypotheses that may be tested and their alternatives.

1. (Two-sided)

H_0: The proportion with the characteristic of interest is the same in both populations; that is, $P_1 = P_2$

H_A: The proportion with the characteristic of interest is not the same in both Populations; $P_1 \neq P_2$

2. (One-sided)

H_0: The proportion with the characteristic of interest in population 1 is less than or the same as the proportion in population 2; $P_1 \leq P_2$
H_A: The proportion with the characteristic of interest is greater in population 1 than in population 2; $P_1 > P_2$

For sufficiently large samples we can test the null hypothesis of the equality of two population proportions by using the normal approximation. Compute

$$Z = \frac{(a/A)-(b/B)}{\sqrt{\hat{P}(1-\hat{P})(1/A + 1/B)}} \rightarrow (3.57)$$

Where: $\hat{P} = ((a+b)/(A+B))$

and compare it for significance with appropriate critical values of the standard normal distribution. The use of the normal approximation is generally considered satisfactory if $a, b, A-a$ and $B-b$ and are all greater than or equal to 5. Alternatively, when sample sizes are sufficiently large, we may test the null hypothesis by means of the chi-square test.

(3.22) Relative Risk and Odds Ratio:

In many fields investigators used the analysis of variance techniques to analyze data that arise from designed experiments, investigations in which at least one variable is manipulated in some way. Designed experiments, of course, are not the only sources of data that are of interest to clinicians and other health sciences professionals. Another important class of scientific investigation that is widely used is the observational study an observational study may be defined simply as an investigation that is not an experiment. The simplest form of observational study is one in which there are only two variables of interest. One of the variables is called the risk factor, or independent variable, and the other variable is referred to as the outcome, or dependent variable, the term risk factor is used to designate a variable that is thought to be related to some outcome variable.

(3.22.1) Types of Observational Studies:

There are two basic types of observational studies, prospective studies and retrospective studies.

A Prospective Study: is an observational study in which two random samples of subjects are selected. One sample consists of subjects who possess the risk factor, and the other sample consists of subjects who do not possess the risk factor. The subjects are followed into the future (that is, they are followed prospectively), and a record is kept on the number of subjects in each sample who, at some point in time, are classifiable into each of the categories of the outcome variable. The data resulting from a prospective study involving two dichotomous variables can be displayed in a 2×2 contingency table that usually provides information regarding the number of subjects with and without the risk factor and the number who did and did not succumb to the disease of interest as well as the frequencies for each combination of categories of the two variables. From the data of a retrospective study we may construct a contingency table with frequencies similar to those that are possible for the data of a prospective study. In general, the prospective study is more expensive to conduct than the retrospective study. The prospective study, however, more closely resembles an experiment.

A Retrospective Study: is the reverse of a prospective study. The samples are selected from those falling into the categories of the outcome variable. The investigator then looks back (that is, takes a retrospective look) at the subjects and determines which ones have (or had) and which ones do not have (or did not have) the risk factor.

Table (3.3): Classification of a Sample of Subjects with Respect to Disease Status and Risk Factor:

Risk factor	Disease status		
	Present	Absent	Total
Present	A	B	a+b
Absent	C	D	c+d
Total	a+c	b+d	N

Source: Wayne, W. Daniel 2009.

Relative Risk:

Relative risk is the ratio of the risk of developing a disease among subjects with the risk factor to the risk of developing the disease among subjects without the risk factor. We represent the relative risk from a prospective study as:

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)} \rightarrow (3.58)$$

where *a*, *b*, *c*, and *d* are as defined in table (3.22.1) and $\widehat{RR}$ indicates that the relative risk is computed from a sample to be used as an estimate of the relative risk, *RR*, for the population from which the sample was drawn. We may construct a confidence interval for *RR:*

$$100(1-\alpha)\%CI = \widehat{RR}^{1\pm(Z_\alpha/\sqrt{x^2})}$$

Where Z_α is the two-sided *z* value corresponding to the chosen confidence coefficient and X^2 is computed by Equation (3.52)

Interpretation of RR:

The value of *RR* may range anywhere between zero and infinity. A value of 1 indicates that there is no association between the status of the risk factor and the status of the dependent variable. In most cases the two possible states of the dependent variable are disease present and disease absent. We interpret an *RR* of 1 to mean that the risk of acquiring the disease is the same for those subjects with the risk factor and those without the risk factor. A value of *RR* greater than 1 indicates that the risk of acquiring the disease is greater among subjects with the risk factor than among subjects without the risk factor. An *RR* value that is less than 1 indicates less risk of acquiring the disease among subjects with the risk factor than among subjects without the risk factor.

Odds Ratio:

When the data to be analyzed come from a retrospective study, relative risk is not a meaningful measure for comparing two groups. As we have seen, a retrospective study is based on a sample of subjects with the disease (cases) and a separate sample of subjects without the disease (controls). We then retrospectively determine the distribution of the risk factor among the cases and controls. Given the results of a retrospective study involving two samples of subjects, cases, and controls, we may display the data in a 2×2 table such as Table (3.22.2) in which subjects are dichotomized with respect to the presence and absence of the risk factor. Note that the column headings in Table (3.22.2) differ from those in Table (3.22.1) to emphasize the fact that the data are from a retrospective study and that the subjects were because they were either cases or controls. When the data from a retrospective study are displayed as in Table (3.22.2) the ratio$[a/(a+b)]$, for example, is not an estimate of the risk of disease for subjects with the risk factor. The appropriate measure for comparing cases and controls in a retrospective study is the odds ratio.

Table (3.4): Subjects of a Retrospective Study Classified According to Status Relative to a Risk Factor and Whether They Are Cases or Controls

Risk factor	Sample		
	Cases	Controls	Total
Present	A	B	a+b
Absent	C	D	c+d
Total	a+c	b+d	N

Source: Wayne, W. Daniel 2009.

The odds for success are the ratio of the probability of success to the probability of failure. By using the definition of odds to define two odds that we can calculate from data displayed as in Table (3.22.2):

1- The odds of being a case (having the disease) to being a control (not having the disease) among subjects with the risk factor is:

$$\frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$$

2- The odds of being a case (having the disease) to being a control (not having the disease) among subjects without the risk factor is:

$$\frac{c/(c+d)}{d/(c+d)} = \frac{c}{d}$$

The odds ratio that we may compute from the data of a retrospective study was defined. We use the symbol $\widehat{OR}$ to indicate that the measure is computed from sample data and used as an estimate of the population odds ratio, *OR*. The estimate of the population odds ratio is:

$$\widehat{OR} = \frac{\mathrm{a/b}}{\mathrm{c/d}} = \frac{\mathrm{ad}}{\mathrm{bc}}$$

where a, b, c, and d are as defined in Table (4). We may construct a confidence interval for *OR* by the following method:

$$100(1-\alpha)\%CI = \widehat{OR}^{1\pm(Z_\alpha/\sqrt{x^2})}$$

Where Z_α is the two-sided z value corresponding to the chosen confidence level and X^2 is computed by Equation (3.52).

Interpretation of the Odds Ratio:

In the case of a rare disease, the population odds ratio provides a good approximation to the population relative risk. Consequently, the sample odds

ratio, being an estimate of the population odds ratio, provides an indirect estimate of the population relative risk in the case of a rare disease.

The odds ratio can assume values between zero and ∞ value of 1 indicates no association between the risk factor and disease status. A value less than 1 indicates reduced odds of the disease among subjects with the risk factor. A value greater than 1 indicates increased odds of having the disease among subjects in whom the risk factor is present.

(3.23) The Mantel–Haenszel Statistic:

Frequently when we are studying the relationship between the status of some disease and the status of some risk factor, we are aware of another variable that may be associated with the disease, with the risk factor, or with both in such a way that the true relationship between the disease status and the risk factor is masked. Such a variable is called a confounding variable. A technique for accomplishing this objective is the Mantel–Haenszel procedure, so called in recognition of the two men who developed it. The procedure allows us to test the null hypothesis that there is no association between status with respect to disease and risk factor status. Initially used only with data from retrospective studies, the Mantel–Haenszel procedure is also appropriate for use with data from prospective studies.

In the application of the Mantel–Haenszel procedure, case and control subjects are assigned to strata corresponding to different values of the confounding variable. The data are then analyzed within individual strata as well as across all strata, (Wayne, W. Daniel 2009).

Assumptions:

Two basic assumptions should be considered when using this procedure.

1- Observations are independent from each other. In practice, this means that each observation comes from a different subject, that the subjects were randomly selected from the population of interest, and that no specific group of subjects is purposefully omitted.
2- All observations are identically distributed. This means that they are obtained in the same way.

Application of the Mantel–Haenszel procedure consists of the following steps:-

1. Form k strata corresponding to the k categories of the confounding variable. Table (3.23.1) shows the data display for the i th stratum.

Table (3.5): Subjects in the ith Stratum of A Confounding Variable Classified According to Status Relative to A Risk Factor and Whether They Are Cases Or Controls

Risk factor	Sample		
	Cases	Controls	Totals
Presents	a_i	b_i	$a_i + b_i$
Absence	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	n_i

Source: Wayne, W. Daniel 2009.

2. For each stratum compute the expected frequency e_i of the upper left-hand cell of table (3.23.1) as follows:

$$e_i = \frac{(a_i + b_i)(a_i + c_i)}{n_i} \rightarrow (3.59)$$

3. For each stratum compute

$$v_i = \frac{(a_i + b_i)(c_i + d_i)(a_i + c_i)(b_i + d_i)}{{n_i}^2(n_i - 1)} \rightarrow (3.60)$$

4. Compute the Mantel–Haenszel test statistic ${\chi^2}_{MH}$, as follows:

$${\chi^2}_{MH} = \frac{\left(\sum_{i=1}^{k} a_i - \sum_{i=1}^{k} e_i\right)^2}{\sum_{i=1}^{k} v_i} \rightarrow (3.61)$$

5. Reject the null hypothesis of no association between disease status and suspected risk factor status in the population if the computed value of ${\chi^2}_{MH}$ is equal to or greater than the critical value of the test statistic, which is the tabulated chi square value for 1 degree of freedom and the chosen level of significance.

(3.23.1) Mantel–Haenszel Estimator of the Common Odds Ratio:

When we have *k* strata of data, each of which may be displayed in a table like Table (3.23.1) the Mantel–Haenszel estimator of the common odds ratio can be computed as follows:

$$\widehat{OR}_{MH} = \frac{\sum_{i=1}^{k} a_i d_i / n_i}{\sum_{i=1}^{k} b_i c_i / n_i} \rightarrow (3.62)$$

The Mantel-Haenszel analysis provides two closely related pieces of information. First, it provides statistical tests of whether the odds ratios are equal (homogeneous) or unequal (heterogeneous) across strata. Second, it provides an estimate of the odds ratio of the exposure variable, adjusted for the strata variable.

(3.23.2) Confidence Limits for the Odds Ratio :

The within-strata odds ratio is computed as in equation (3.62). The test-based confidence limits for a $100(1-\alpha)\%$ confidence interval are given by:

$$OR_{MH,lower} = exp\left[\left(1 - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{\chi^2{}_{MH}}}\right)\ln(OR_{MH})\right]$$

$$OR_{MH,upper} = exp\left[\left(1 + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{\chi^2{}_{MH}}}\right)\ln(OR_{MH})\right]$$

The Mantel-Haenszel chi-square value tests the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity.

(3.24) Some Important Concepts in Mantel-Haenszel Procedure:

Delta(δ):

This value is added to each cell count when zeroes are present in the table. If there are no zeroes in the table, then this value is ignored. This option lets you analyze data with zero counts. The traditional value is 0.5. Recent simulation studies have indicated that 0.25 produces better results under certain situations.

(3.24.1) Corrected Odds Ratio:

This odds ratio is computed using the formula:

$$\acute{OR} = \frac{(a+\delta)(d+\delta)}{(c+\delta)(b+\delta)} \rightarrow (3.63)$$

where δ is the Delta value that was entered (usually, 0.5 or 0.25). This odds ratio is defined when one or more cell counts are zero.

(3.24.2) Lower and Upper 100(1-alpha) % C.L. :

The odds ratio confidence limits are calculated from those based on the Log Odds Ratio using the following procedure.

1. Compute the corrected odds ratio $\acute{OR}$ using the formula above.
2. Compute the logarithm of the odds ratio using:

$$\grave{L} = \ln(\acute{OR})$$

3. Compute the standard error of $\grave{L}$ using:

$$S_{\grave{L}} = \sqrt{\frac{1}{(a+\delta)} + \frac{1}{(b+\delta)} + \frac{1}{(c+\delta)} + \frac{1}{(d+\delta)}} \rightarrow (3.64)$$

4. Compute the $100(1-\alpha)\%$ confidence limits for L using the fact that $\grave{L}$ is approximately normally distributed for large samples:

$$\grave{L} \pm Z_{\frac{\alpha}{2}} S_{\grave{L}}$$

where $Z_{\frac{\alpha}{2}}$ is the appropriate value from the standard normal distribution.

5. Transform the above confidence limits back to the original scale using:

$$OR_{lower} = e^{\grave{L} - Z_{\frac{\alpha}{2}} S_{\grave{L}}}$$
$$OR_{upper} = e^{\grave{L} + Z_{\frac{\alpha}{2}} S_{\grave{L}}}$$

(3.24.3) Proportion Exposed and Proportion Diseased:

Proportion exposed is the overall proportion of those in the table that were exposed to the risk factor. The calculation is:

$$prportion\ exposed = \left(\frac{a_i + b_i}{n_i}\right) \rightarrow (3.65)$$

Proportion diseased is the overall proportion of those in the table that were diseased. The calculation is:

$$proportion\ diseased = \left(\frac{a_i + c_i}{n_i}\right) \rightarrow (3.66)$$

(3.24.4) Mantel-Haenszel with Continuity Correction (MHC.C.):

By using this procedure the confidence limits and hypothesis test with continuity correction can be clarified. Generally speaking, the continuity correction is used to provide a closer approximation to the exact conditional test in which all marginal totals are assumed to be fixed. Bennett and Kaneshiro (1974) suggested that use of a continuity correction in the Mantel-Haenszel test is unnecessary for small samples, (Li, Shou-Hua et al. 1979). The Mantel-Haenszel chi-square value tests the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The Heterogeneity Test is different from which does not test that the odds ratios are equal to one, just equal to each other.

$$\chi^2{}_{MHC.C.} = \frac{\left(\left|\sum_{i=1}^{k} a_i - \sum_{i=1}^{k} E(a_i) - \frac{1}{2}\right|\right)^2}{\sum_{I=1}^{K} v_i} \rightarrow (3.67)$$

Where K is the number of strata and

$$E(a_i) = \frac{(a_i + b_i)(a_i + c_i)}{n_i}$$

$$v_i = \frac{(a_i + b_i)(a_i + c_i)(c_i + d_i)(b_i + d_i)}{{n_i}^2(n_i - 1)}$$

(3.24.5) Confidence Limits for the Odds Ratio:

The within-strata odds ratio is computed as follows:

$$OR_{MH} = \frac{\sum_{i=1}^{k} a_i d_i / n_i}{\sum_{i=1}^{k} b_i c_i / n_i}$$

The test-based confidence limits for a $100(1-\alpha)\%$confidence interval are given by:

$$OR_{MHC,lower} = exp\left[\left(1 - \frac{Z_{\frac{\alpha}{2}}}{\sqrt{\chi^2{}_{MHC.C.}}}\right)\ln(OR_{MH})\right] \rightarrow (3.68)$$

$$OR_{MHC,upper} = exp\left[\left(1 + \frac{Z_{\frac{\alpha}{2}}}{\sqrt{\chi^2{}_{MHC.C.}}}\right)\ln(OR_{MH})\right] \rightarrow (3.69)$$

(3.25) Woolf procedure:

Woolf developed a new procedure to illustrated confidence limits and hypothesis test. Recent studies have cast doubt on the usefulness of Woolf's tests, but they are provided anyway for completeness. Woolf's chi-square statistic tests the hypothesis that all odds ratios are equal to one.

Confidence Limits for the Odds Ratio:

The within-strata odds ratio is computed as follows:

$$OR_w = exp\left[\frac{\sum_{i=1}^{k} {v_i}^{-1} \ln OR_i}{\sum_{i=1}^{k} {v_i}^{-1}}\right] \rightarrow (3.70)$$

$$v_i = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \rightarrow (3.71)$$

$$W = \sum_{i=1}^{k} {v_i}^{-1} \rightarrow (3.72)$$

$$OR_{w,lower} = OR_w exp\left(\frac{-Z_{\alpha/2}}{\sqrt{w}}\right) \rightarrow (3.73)$$

$$OR_{w,upper} = OR_w exp\left(\frac{Z_{\alpha/2}}{\sqrt{w}}\right) \rightarrow (3.74)$$

Hypothesis Test:

Woolf's chi-square statistic tests the hypothesis that all odds ratios are equal to one. The formula used for this test is:

$${\chi^2}_w = W(\ln OR_w)^2 \rightarrow (3.75)$$

This is a chi-square test with one degree of freedom. The probability level provides the upper tail probability of the test. Hence, when this value is less the alpha level (say 0.05), reject the null hypothesis that all odds ratios are equal to one.

(3.26) Heterogeneity Test:

This test focuses on a hypothesis test developed by Woolf for testing the more general hypothesis that all odds ratios are equal, but not necessarily equal to one.

Hypothesis Test:

Woolf's chi-square statistic tests the hypothesis that all odds ratios are equal. The formula used for this test is:

$${\chi^2}_{wh} = \sum_{i=1}^{k} {v_i}^{-1}(\ln OR_i - \ln OR)^2 \rightarrow (3.76)$$

This is a chi-square test with K-1 degrees of freedom. The probability level provides the upper tail probability of the test. Hence, when this value is less that the desired alpha level, reject the null hypothesis that all odds ratios are equal to one, (NCSS, LLC. 2016).

Chapter 4

Data Analysis & Application

Chapter Four

(4.1) preface:

This research study was planned to study the epidemiology of prostate cancer among Sudanese men. After discussion with the oncologists, an interview schedule was prepared. The structured questionnaire contained the questions regarding all the variables that could be related to prostate cancer risk and should be studied for the population of Sudan. For the design of the research, a case-control study was planned. The case-control study is a primary tool for the study of factors related to the disease incidence. The data for patients were collected from the National Center for Radiotherapy and Nuclear Medicine in Khartoum state, target population is all men in Khartoum state during the period of survey. The data were checked to confirm whether the patients attending the hospital were diagnosed with prostate cancer cases. The same structured questionnaire was used to interview the cases and controls. The questionnaire was filled in personally from the respondents.

In this chapter the descriptive analysis of the cancer data was performed by NCSS11. All the data were coded, organized and entered into computer for advanced analysis. To conduct this research on prostate cancer, the outcome variable was binary. Its value (1) represented the patients (cases) and (0) represented healthy men (controls). The data were collected for 150 cases and 100 controls for all variables under study. These variables were; Age, Occupation, State , Marital status, Family history, Animal fat, Fruits and green vegetables, Overweight, Cholesterol, Blood pressure, Prostate medications, Alcohol, Smoking, Prostate Specific Antigen (PSA), and developing one or more of these diseases: “Syphilis, gonorrhea, chronic prosatitis, and prostate enlargement”. All these variables (the risk factors) classified into two categories denoted by (1) and (0) to represent presence and absence of the risk factors. Except the variable “state” classified into four categories; (1) represented Darfur and Kurdufan states, (2) represented states of (Formerly Central Region), (3) represented the states of Northern

and Eastern Sudan and (4) represented Khartoum state. The researcher focused on the last state because it is the capital of Sudan and has a high population density of up to eight million people. All variables that were statistically associated with prostate cancer incidence were indentified.

Multiple-logistic regression model was fitted to obtained independent estimates of the risk of prostate cancer. Modeling started with a constant (with no variable) followed by sequential selection according to their statistical importance, the outcome variable was binary. A comparison between chi-square test and Mantel-Haenszel test was performed to assess the risk factors most related to the prostate cancer incidence. Odds ratios and their 95 percent confidence intervals were determined.

(4.2) Data Collection and Sample Size:

This research is a retrospective study using data from National Center for Radiotherapy and Nuclear Medicine which covered the characteristics of men diagnosed with prostate cancer. The researcher collected data by questionnaire through personal interviews and from their medical history during one year (2015 – 2016), daily for 317 days {except Fridays}. The sample size was 250 of whom150 volunteers were cases and 100 were controls. Selection of the cases and controls was performed on the basis of the outcome. The control men selected randomly without prostate cancer.

The data were collected using same questionnaire including the same information. The sample size was calculated by the formula:

$$n = \frac{Z^2_{\frac{\alpha}{2}} pq}{d^2}$$

where: $Z_{\frac{\alpha}{2}} = 1.96$, $p = 0.38$, $q = 0.62$, $d = 0.06$, so;

$$n = \frac{(1.96)^2(0.38)(0.62)}{(0.06)^2} = 251.4 \cong 250$$

The simple random sample was selected with size 250 men. Then the size of patients sample was 150, which determined by the formula:

$$\grave{n} = \frac{1}{\frac{1}{n} + \frac{1}{N}}$$

n : Originally calculated size (250).
$\grave{n}$: Adjusted sample size (patients sample).
N: Population size in Khartoum state is 375.

(4.3) Descriptive Statistics:

In this section descriptive statistics is done to explain the main characteristics of the study population through the sample. Cross tabulation between the disease status and risk factors was made to illustrate: counts, percentages and expected values (assuming the independence between two variables) in each cell. The preliminary analysis of the data set was carried out. A comparison between two groups (cases and controls) of data was made for all variables. The results shown below:

Table (4.1)

The Sample Distribution According to Age

Age \ Diagnostic		Without disease	With disease	Total
Less than 50	Counts	91	5	96
	Percentages	36.40%	2.00%	38.40%
	Expected counts	38.4	57.6	96.0
Greater than or equal 50	Counts	9	145	154
	Percentages	3.60%	58.00%	61.60%
	Expected counts	61.6	92.4	154.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

The variable “age” was categorized into two categories, less than 50 and greater than or equal 50. Table (4.3.1) showed the total number of controls

and cases was 100(40%) and 150(60%), respectively. The total percentage of men under 50 was 38.4%, concluded 36.4% without prostate cancer and just 2% with disease. Also the total percent of the men above 50 was 61.6%; concluded 58% with disease was and 3.6% without. The expected value of each cell has large difference than observed value that means no agreement between two values.

Table (4.2)

The Sample Distribution According to Occupation

Diagnostic / Occupation		Without disease	With disease	Total
Group1	Counts	89	87	176
	Percentages	35.60%	34.80%	70.40%
	Expected counts	70.4	105.6	176.0
Group2	Counts	11	63	74
	Percentages	4.40%	25.20%	29.60%
	Expected counts	29.6	44.4	74.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

In this study the variable "occupation" has two categories: group1 represents males who have jobs depend on prolonged seating, and group2 represents males with jobs depend on physical activities. Table (4.3.2) illustrated the total percentages of group1 was 70.4% included 35.6% without prostate cancer and 34.8% with. 29.6% represented total percent of group2, with 25.2% of them suffering from disease and 4.4% don't. There was a little agreement between observed and expected value.

Table (4.3.3) illustrated that the percentage of cases was 60% of whom16.8% live in the states of the former central region, and 15.6% represented percentage of Darfur and kurdufan states and followed by the Khartoum state with 15.2%. The lowest percentage of the disease incidence was 12.4% in the states of northern and eastern Sudan. The percentage of controls was 40% of whom 26.4% from Khartoum state, followed by states of formerly central region with 8.8%. Both (Darfur and kurdufan states) and (the states of northern and eastern Sudan) had same percentage in terms of

the proportion of healthy people (without cancerous prostate). Most of the study participants were residents of Khartoum State with 41.6%, and the least participants were residents of the states of northern and eastern Sudan with 14.8%. The participation of Darfur and kurdufan states was 18% and 25.6% for states of the former central region. A weak agreement between observed and expected values can be observed.

Table (4.3)

The Sample Distribution According to State

Diagnostic / State		Without disease	With disease	Total
Darfur and kurdufan states	Counts	6	39	45
	Percent	2.40%	15.60%	18.00%
	Exp.value	18.0	27.0	45.0
states of (previously central region)	Counts	22	42	64
	Percent	8.80%	16.80%	25.60%
	Exp.value	25.6	38.4	64.0
The states of northren and eastern sudan	Counts	6	31	37
	Percent	2.40%	12.40%	14.80%
	Exp.value	14.8	22.2	37.0
khartoum state	Counts	66	38	104
	Percent	26.40%	15.20%	41.60%
	Exp.value	41.6	62.4	104.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.4): The Sample Distribution According to Marital Status:

Diagnostic / Marital status		Without disease	With disease	Total
Married	Counts	40	146	186
	Percentages	16.00%	58.40%	74.40%
	Expected counts	74.4	111.6	186.0
Single	Counts	60	4	64
	Percentages	24.00%	1.60%	25.60%
	Expected counts	25.6	38.4	64.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.4) clarified that there were 60% of men suffering from prostate cancer of whom 58.4% married men and 1.6% were not. The percentage of men without prostate cancer was 40% of whom 16% were married and 24% were single. In this study the total percentage of married and single men was 74.4% and 25.6% respectively. From this table we observed difference between real values and expected values.

Table (4.5)

The Sample Distribution According to Family History:

Diagnostic / Family history		Without disease	With disease	Total
No	Counts	81	94	175
	Percentages	32.40%	37.60%	70.00%
	Expected counts	70.0	105.0	175.0
Yes	Counts	19	56	75
	Percentages	7.60%	22.40%	30.00%
	Expected counts	30.0	45.0	75.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

In this study the percentage of cases with relatives infected with same disease was 22.4% and 37.6% represented the cases with healthy relatives, reported in table (4.3.5). While the percentage of controls with healthy relatives was 32.4%% and just7.6% represented controls with infected relatives. The total percentage of cases and controls were 60% and 40% respectively. The rates of the participants with infected relatives and without were 30% and 70% respectively. Convergence between the real values and expected values can be observed.

Table (4.6)

The Sample Distribution According to Eating Red Meats & Animal Fat Regularly:

Diagnostic / Animal fat		Without disease	With disease	Total
No	Counts	38	26	64
	Percentages	15.20%	10.40%	25.60%
	Expected counts	25.6	38.4	64.0
Yes	Counts	62	124	186
	Percentages	24.80%	49.60%	74.40%
	Expected counts	74.4	111.6	186.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.6) illustrated that 60% and 40% represented the cases and controls respectively. The percentage of infected men who relied on diet rich in animal fats was 49.6%, and only 10.4% were followed balanced diet. Whilst the proportion of healthy men (men who do not have prostate cancer) who were ate red meat and animal fat regularly was 24.8% and 15.2% were dependent on a healthy diet. In this study, participants who followed a good diet were 25.6%, while 74.4% had a diet rich in animal fats. A clear convergence between observable and predictive values can be noted in this table.

Table (4.3.7) clarified that 60% and 40% represented the cases and controls respectively. The percentage of infected men who relied on diet rich in fruits and vegetables was 34.4% and 25.6% did not prefer to eat these food ingredients regularly. While the proportions of healthy men (men who do not have prostate cancer) who were ate green vegetables and fruits regularly was 9.6% and 30.4% were not. In this study, volunteers who followed diet rich in fruits and vegetables were 44%%, while 56% were not.

Table (4.7)

The Sample Distribution According to Eating Green Vegetables & Fruits Regularly:

Diagnostic / Fruits & vegetables		Without disease	With disease	Total
No	Counts	76	64	140
	Percentages	30.40%	25.60%	56.00%
	Expected counts	56.0	84.0	140.0
Yes	Counts	24	86	110
	Percentages	9.60%	34.40%	44.00%
	Expected counts	44.0	66.0	110.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.8)

The Sample Distribution According to Overweight:

Diagnostic / Overweight		Without disease	With disease	Total
No	Counts	81	99	180
	Percentages	32.40%	39.60%	72.00%
	Expected counts	72.0	108.0	180.0
Yes	Counts	19	51	70
	Percentages	7.60%	20.40%	28.00%
	Expected counts	28.0	42.0	70.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.8) showed the total percentage of participants who were suffered from overweight was 28% and 72% referred to infected men who had normal weight. 20.4% represented men with prostate cancer who suffered from obesity, and 39.6% had normal weight. The proportion of controls (men without cancer), who were suffered from obesity was 7.6% and 32.4% did not. We noted a little agreement between observed and expected values.

Table (4.9)

The Sample Distribution According to High Cholesterol:

Diagnostic / Cholesterol		Without disease	With disease	Total
No	Counts	97	109	206
	Percentages	38.80%	43.60%	82.40%
	Expected counts	82.4	123.6	206.0
Yes	Counts	3	41	44
	Percentages	1.20%	16.40%	17.60%
	Expected counts	17.6	26.4	44.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.9) demonstrated that the percentage of cases was 60% of whom 16.4% were suffered from high cholesterol and 43.6% were not. 40% of participants represented controls, of whom 38.8% had normal cholesterol and only 1.2% had not. 17.6% represented total percentage of volunteers who suffered from high cholesterol and 82.4% did not suffer. No large difference between observed and expected counts.

Table (4.3.10) illustrated that the percentage of cases was 60% of whom 17.2% were suffered from high blood pressure and 42.8% were not. 40% of participants represented controls, of whom 38% had normal blood pressure and only 2% had not. 19.2% represented total percentage of volunteers who

suffered from high blood pressure and 80.8% did not suffer. No large difference between observed and expected counts.

Table (4.10)

The Sample Distribution According to High Blood Pressure:

Diagnostic / Blood pressure		Without disease	With disease	Total
No	Counts	95	107	202
	Percentages	38.00%	42.80%	80.80%
	Expected counts	80.8	121.2	202.0
Yes	Counts	5	43	48
	Percentages	2.00%	17.20%	19.20%
	Expected counts	19.2	28.8	48.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.11)

The Sample Distribution According to Intake of Prostate Medications:

Diagnostic / Prostate med.		Without disease	With disease	Total
No	Counts	5	52	57
	Percentages	2.00%	20.80%	22.80%
	Expected counts	22.8	34.2	57.0
Yes	Counts	95	98	193
	Percentages	38.00%	39.20%	77.20%
	Expected counts	77.2	115.8	193.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.11) showed that 60% and 40% represented the cases and controls respectively. The percentage of men with prostate cancer who took medications to treat prostate diseases was 39.2%, and 20.8% had not take. Whilst the proportion of healthy men (men without prostate cancer) who took prostate medications was 38% and only 2% had not take. In this study, volunteers who were suffered from prostate diseases in their live and took the treatments were 77.2%, while 22.8% did not take. Disagreement between observable and predictive values can be noted in this table

Table (4.12)

The Sample Distribution According to Alcohol Consumption

Diagnostic / Alcohol		Without disease	With disease	Total
No	Counts	93	90	183
	Percentages	37.20%	36.00%	73.20%
	Expected counts	73.2	109.8	183.0
Yes	Counts	7	60	67
	Percentages	2.80%	24.00%	26.80%
	Expected counts	26.8	40.2	67.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.12) clarified that 60% and 40% represented the cases and controls respectively. The percentage of men with prostate cancer who were consumed alcohol was 24%, and 36% did not drink. Whilst the proportion of healthy men (men without prostate cancer) who did not drink alcohol was 37.2% and only 2.8% were consumed alcohol that did not affect on the prostate. In this study, the total proportion of the volunteers who were consumed alcohol was 26.8%, while 73.2% did not. A little agreement between observable and predictive values can be observed in this table.

Table (4.3.13) demonstrated that the percentage of cases was 60% of whom 32.8% were smokers and 27.2% were not. 40% of participants represented controls, of whom 22.4% were smokers and only 17.6% did not.

The percentage 55.2% represented the total percentage of volunteers who were smokers and 44.8% did not. In this table the large convergence between observed and expected counts can be found.

Table (4.13)

The Sample Distribution According to Smoking

Smoking \ Diagnostic		Without disease	With disease	Total
No	Counts	44	68	112
	Percentages	17.60%	27.20%	44.80%
	Expected counts	44.8	67.2	112.0
Yes	Counts	56	82	138
	Percentages	22.40%	32.80%	55.20%
	Expected counts	55.2	82.8	138.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.14)

The Sample Distribution According to Developing One or More of These Diseases: "Syphilis, Gonorrhea, Chronic Prostatitis and Prostate Enlargement":

Diseases \ Diagnostic		Without disease	With disease	Total
No	Counts	94	50	144
	Percentages	37.60%	20.00%	57.60%
	Expected counts	57.6	86.4	144.0
Yes	Counts	6	100	106
	Percentages	2.40%	40.00%	42.40%
	Expected counts	42.4	63.6	106.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.14) clarified that the overall proportion of participants who developed one or more of these diseases: "syphilis, gonorrhea, chronic prostatitis and prostate enlargement" was 42.4% and 57.6% represented men who did not suffer from these diseases. 40% represented men with prostate cancer who developed from the above diseases, and 20% had no one of these diseases. The percentage of controls (men without cancer), who were suffered from the above diseases was 2.4% and 37.6% did not suffer. We noted that there was large difference between observed and expected values.

Table (4.15)

The Sample Distribution According to PSA:

Diagnostic / PSA		Without disease	With disease	Total
Normal	Counts	94	3	97
	Percentages	37.60%	1.20%	38.80%
	Expected counts	38.8	58.2	97.0
Abnormal	Counts	6	147	153
	Percentages	2.40%	58.80%	61.20%
	Expected counts	61.2	91.8	153.0
Total		100(40%)	150(60%)	250(100%)

Source: The researcher output using NCSS software.

Table (4.3.15) illustrated that the percentage of cases was 60% of whom 58.8% were suffered from high prostate specific-antigen (PSA) and only 1.2% had normal PSA. While 40% of volunteers represented controls, of whom 37.6% had normal PSA and only 2.4% were suffered from high PSA. 61.2% represented overall percentage of volunteers who suffered from high PSA and 38.8% did not suffer. No agreement between observed and expected counts.

(4.4) Logistic Regression Analysis:

For this research on prostate cancer, the outcome variable was binary. Its value (1) represented (cases) and (0) represented (controls). The data were collected for 150 cases and 100 controls for all variables in the questionnaire. The variables or the risk factors most related to the incidence of prostate cancer were mentioned. All variables (the risk factors) classified into two categories denoted by (1) and (0) to represent presence and absence of the risk factors. Only, state variable was classified into four categories.

In this section the multiple logistic regression models were used to build an accurate statistical model that describes the relationship between the incidence of prostate cancer and the risk factors and to assess the variables most related to the disease. Odds ratios confidence intervals were calculated. The variables which significant associated to outcome variable at (0.05) were: age, PSA, state and alcohol. The method of forward selection was used to select the variables according to their importance, in this way the saturated model was built at five steps. The following tables showed the explained results of using this statistical technique:

Table (4.16)

Run Summary

Item	Value	Item	Value
Y Variable	Diagnostic	Rows Processed	250
Reference Value	without disease	Rows Used	250
Number of Y-Values	2	Rows for Validation	0
Frequency Variable	None	Rows X's Missing	0
Numeric X Variables	0	Rows Freq Miss. or 0	0
Categorical X Variables	15	Rows Prediction Only	0
Final Log Likelihood	-20.05139	Sum of Frequencies	250
Model R^2	0.88083	Likelihood Iterations	9
Actual Convergence	7.453633E-09	Maximum Iterations	20
Target Convergence	1E-06	Completion Status	Normal Completion
Model D.F.	7		
Priors	Ni/N		
Subset Selection Method	Forward Selection		

Source: The researcher output using NCSS software.

Table (4.4.1) specifies the independent variable's name (Diagnostic), and reference value was (without disease); this option specified a reference value for the dependent variable, it was the outcome for which no regression equation was generated, this value could be text or numeric. Number of y-values referred to number of categories for outcome variables, it had two categories (0) and (1) represented controls and cases respectively.
Frequency variable specified an optional frequency (count) variable. This variable contained integer that represented the number of observations (or frequency) associated with each observation. If left blank, each observation has a frequency of one. This is especially useful when the data are already tabulated and you want to enter the counts. All the variables under study were categorized into two or more categories. The prior probabilities were estimates of the probabilities that a new individual exhibits each possible outcome. We used the choice of Ni/N (Y-Value Proportions) for the prior probabilities as estimated by the Y-value proportions of the data.

Forward selection method was used to select the best subset from independent variables (X's) with maximum iteration 20. The selection stops at five steps. The final log likelihood is equal (-20.05). The R^2 values tell us approximately how much variation in the outcome was explained by the model (88%), this implies that 88% percent of variation in Y caused by the independent variables. The Target Convergence (0.000001) represented the amount that was used to stop the iterative fitting of the maximum likelihood algorithm. If the Actual Convergence amount was larger than the Target amount, the algorithm ended before converging, and care must be taken in using any of the results. The Likelihood Iterations are the number of iterations necessary to solve the likelihood equations. In this case nine iterations were necessary. The degree of freedom to the model was 7

Table (4.17)
Dependent Variable (Y) Summary

Y Diagnostic	count	Y proportion	R^2 (Y vs. Pred. probability)	Percent correctly classified
Without disease	100	0.40	0.92602	97.00
With disease	150	0.60	0.92602	99.33
Total	250	1.00		98.40

Source: The researcher output using NCSS software.

Table (4.4.2) with title "variable summary" described the outcome

variable; it summarized the number of individuals with and without disease. 250 individuals were collected of whom 100 without disease and 150 with disease, with proportion 40% and 60% respectively. And R^2 for the Y versus predicted probability was 0.93, and the percentage 97% demonstrated the correctly classified of individuals without disease whoever, 99.3% correctly classified as with disease, with percentage of total for all 98.4%.

Table (4.18)
Subset Selection Summary

Subset Selection Method = Forward Selection

No. Term	No. X's	Log Likelihood	R² value	R² change	Entered
1	1	-168.25292	0.00000	0.00000	Intercept
2	2	-38.69427	0.77002	0.77002	PSA
3	3	-24.90294	0.85199	0.08197	Age
4	6	-22.15279	0.86834	0.01635	State
5	7	-20.05139	0.88083	0.01249	Alcohol

Source: The researcher output using NCSS software.

Table (4.4.3) clarified that the forward selection method was used to choice the best subset variables from the independent variables (X's). A five steps criterion conducted to get the best variables with the value of log likelihood, first step for intercept so the R^2 (usually use R^2 to determinant the important variable) was zero so, there was no variables, with the (-168.25292) log likelihood, in the second step, (PSA) entered to the model with $R^2 = 0.77$ and log likelihood (-38.69427), in third step age entered so, R^2 was changed by 0.08 ($R^2 = 0.85$) and the log likelihood increased to (-24.90294), fourth step state was entered also, R^2 was changed by 0.016 ($R^2 = 0.868$) also log likelihood decreased to (-22.15279), the last step (step five) the variable alcohol consumption was entered, also R^2 was changed by 0.012 ($R^2 = 0.88$) also log likelihood increased to (-20.05139) so, the ranking of the important variables as in above.

In this study alcohol consumption, state=2 (the states of northern and eastern Sudan) and state=1 (Darfur and kurdufan states) were insignificant, so they had no effect in the model as shown in table (4.4.4). Also, this study showed the significant variables were age, state=3 and PSA with p-value 0.00025, 0.03426 and 0.00031 respectively. The prostate cancer had increased in men over the age of 50 years with ($\beta = 0.16762$) and odds ratio ($OR = 1.2$), which means that log of odds for incidence of prostate

cancer was greater in men over 50 years (1.2) times than men less than 50 years

Table (4.19)
Coefficient Significance Tests:

Independent Variable X	Regression Coefficiens β(i)	Standard Error sb(i)	Wald Z-value H0: β=0	Wald P-Value	Lower 95% confidence limit	Odds Ratio Exp(b(i))	Upper 95% confidence limit
Intercept	-10.71275	2.81841	-3.801	0.00014	-16.23673	0.00002	-5.18877
Age	0.16762	0.04580	3.660	0.00025	0.07785	1.18249	0.25739
(State=1)	1.29516	1.12620	1.150	0.25013	-0.91216	3.65158	3.50248
(State=2)	1.74882	1.43864	1.216	0.22413	-1.07086	5.74784	4.56851
(State=3)	4.35512	2.05720	2.117	0.03426	0.32309	77.87620	8.38715
(PSA=1)	4.62114	1.28241	3.603	0.00031	2.10766	101.60999	7.13462
(Alcohol =1)	-2.28197	1.28458	-1.776	0.07566	-4.79970	0.10208	0.23577

Source: The researcher output using NCSS software.

Abnormal PSA increased the risk of the disease with very large odds ratio (101) and ($\beta = 4.62114$), that means PSA is most important variable in this study.

The state variable also had important role in this study, this variable consists of 4 categories (Khartoum state was the reference group). States of (formerly central region) "State=3" had large odds ratio equal to (77.9), that means men in States of (the former central region) had greater incidence rate (77.9) times than men lived in Khartoum state, and (β =4.35512).

Table (4.20)

Classification Table

Actual	Estimated		Total
	Without disease	With disease	
Without disease	97	3	100
With disease	1	149	150
Total	98	152	250

Source: The researcher output using NCSS software.

Table (4.4.5) aimed to know the difference between the actual and estimated values were conducted by the model. There were 100 individuals actually without disease, while the number of the individuals without disease estimated by the model was (98). On the other hand, the model estimated (152) people to be diagnosed with the disease. There were (150) diagnosed with disease. So, the classification percentage of the model was 98.4%.

Estimated Logistic Regression Model(s) in Reading Form:

Model for Logit (diagnostic) = XB when diagnostic = with disease
-10.71 + 0.17 * age + 1.30 * (state=1) + 1.75 * (state=2) + 4.36 * (state=3) + 4.62 * (PSA=1) - 2.28 * (alcohol consumption=1)

Each model estimates XB (where Logit(Y) = XB) for a specific Y outcome. To calculate the Y-value probabilities when there are only 2 outcomes, transformation of the logit can be used as:
Prob(Y = outcome) = 1/ (1+Exp (-XB))
or Prob(Y ≠ outcome) = Exp (-XB)/ (1+Exp (-XB)).

(4.5) Chi-Square Test:

The chi-square distribution is the most frequently employed statistical technique for the analysis of count or frequency data. It was used to in this section to test independence between the risk factors and the outcome variable. Also the Mantel-Haenszel test was used to achieve one of the goals of the study was that to study the relationship between the status of the disease and the status of some risk factor by using another variable (a confounding variable) that may be associated with the (disease, risk factor or both) used to clarify the true relation between two variables.

The main objective of this section was to compare between most tests used to analyze categorical data such as contingency tables or stratified tables. Also this section aimed to assess the variables agreed by the Mantel-Haenszel test and the Chi-square test that affected on the incidence of prostate cancer. These tests had been used to determine the variables most related to outcome variable (the incidence of prostate cancer). If chi-square assumptions had not been met, the Fisher exact test was used. Study subjects consisted of patients who were treated for prostate cancer during one year (2015 – 2016). The researcher collected the data from patients daily for 317 days. The data consist of 15 independent variables: age, occupation, the state, marital status(maritstat), family history(famhist), eating red meats and animal fat regularly(animfat), eating green vegetables and fruits regularly(greenveg), suffering from overweight(weight), high

cholesterol(cholesterol), high blood pressure(bloodpres), ingestion of prostate medication (prostatmed), alcohol consumption(alcohol), smoking, developing one or more of these diseases: "syphilis, gonorrhea, chronic prostatitis, prostate enlargement " (diseases) and prostate specific antigen (PSA). The case–control study was carried out in the National Center for Radiotherapy and Nuclear Medicine in Khartoum state, Sudan. The sample size was 250 individuals; 150 cases (with prostate cancer) and 100 were controls (without prostate cancer). The outcome variable is (Diagnostic) represented the incidence of prostate cancer (1: Yes and 0: No)**.** The risk factors which significantly associated with the outcome variable were identified by using NCSS11. Both tests agreed that these variables: (PSA, diseases, alcohol, weight, greenveg, animfat and maritstat) were significantly related to the disease. The results as shown below:

(4.5.1) Results Of Chi-Square Test:

The null hypothesis states that the risk factor and outcome variable (diagnostic) are independent, and alternative hypothesis states the two variables were dependent. When p-value less than some specific α (0.05) the null hypothesis would be rejected. Large p-value (greater than α) indicated that there was evidence to accept the null hypothesis:

Chi-square test demonstrated that all variables were significantly associated with the outcome variable (diagnostic), except the smoking variable was insignificant shown in table (4.5.1). In this table risk factors were arranged by their importance. The most important independent variable was PSA (χ^2 = 213.87, Yates's = 210.00) with p-value (0.00000). In case that the 2×2 contingency tables had cell counts less than 5, the Fisher exact test had been used to prove that there was association between the independent variables and the prostate cancer incidence instead of χ^2 test. So; the variables of PSA, marital status and cholesterol are associated to the variable (diagnostic) with Fisher p-value (0.00000). These variables were illustrated in the table with the sign (†) associated with the value of χ^2. The second risk factor was age, there was strong association between age and outcome variable (χ^2 = 194.94, Yates's = 191.25) with p-value (0.00000). When the age became greater than or equal 50, this indicated that the individuals had strong risk factor for prostate cancer incidence. The results of this test showed that there were no association between smoking and the disease (χ^2 = 0.0431, Yates's = 0.0061) with p-value (0.938), so the test was insignificant.

Table (4.21)
Chi-Square Test of Independence:

Risk factor	Person's chi-square (χ^2 value)	Yates's Correction (χ^2 value)	Df	Fisher exact (Prob. level)	Prob. Level	Reject H_0 at α = 0.05?
PSA	213.8670†	210.0101	1	0.00000	0.00000	Yes
Age	194.9433	191.2548	1	0.00000	0.00000	Yes
Maritstat	103.5506†	100.5623	1	0.00000	0.00000	Yes
Diseases	90.4197	87.9527	1	0.00000	0.00000	Yes
State	46.7504	*	3	*	0.00000	Yes
Alcohol	33.3068	31.6459	1	0.00000	0.00000	Yes
Prostatmed	30.0011	28.3393	1	0.00000	0.00000	Yes
Occupation	27.6701	26.2024	1	0.00000	0.00000	Yes
Greenveg	27.0563	25.7204	1	0.00000	0.00000	Yes
Cholesterol	24.4971 †	22.8479	1	0.00000	0.00000	Yes
Bloodpress	21.6627	20.1640	1	0.00000	0.00000	Yes
Animalfat	13.4549	12.3917	1	0.00035	0.00043	Yes
Famhist	9.6032	8.7500	1	0.00196	0.00310	Yes
Weight	6.6964	5.9730	1	0.00995	0.01453	Yes
Smoking	0.0431	0.0061	1	0.89692	0.93792	No

Source: The researcher output using NCSS software.

* Test computed only for 2×2 tables.

† Warning: At least one cell had a value less than 5.

Chi-square test was accurate and suitable for most variables tested and which were classified into two categories or more. The variable "state" had 4 categories, this explains that why the Yates's correction and Fisher exact were inapplicable. So; the chi square test was used to prove the association to the disease. This variable is illustrated in the table with a sign (*)

(4.6) Mantel-Haenszel Tests:

In this section risk factors were identified and the odds ratios were calculated. A comparison between p-value and α was made, to test the null hypothesis which states there is no association between the risk factor and the outcome variable. The results shown below:

Table (4.22)
Mantel-Haenszel Test

Risk factor	χ^2-value	df	Prob. Level	Estimated odds ratio	Lower 95.0% C.L.	Upper 95.0% C.L.	Reject H_0 at α = 0.05?
PSA	121.11	1	0.000000	173.0926	69.1265	433.4235	Yes
Age	54.68	1	0.000000	56.1608	19.3090	163.3455	Yes
Maritalstat	26.25	1	0.000000	0.0197	0.0044	0.0886	Yes
Diseases	13.77	1	0.000207	6.5452	2.4261	17.6574	Yes
Alcohol	10.05	1	0.001523	8.0659	2.2188	29.3216	Yes
Weight	5.33	1	0.021021	3.6003	1.2129	10.6870	Yes
Greenveg	4.75	1	0.029287	3.8013	1.1440	12.6305	Yes
Animfat	4.05	1	0.044048	4.0880	1.0382	16.0963	Yes
Occupation	3.79	1	0.051448	4.2533	0.9909	18.2556	No
Cholesterol	2.77	1	0.096053	3.9273	0.7843	19.6670	No
Prostatmed	1.69	1	0.194113	0.5032	0.1784	1.4189	No
Familyhist	0.71	1	0.397929	1.7424	0.4809	6.3139	No
Smoking	0.30	1	0.586767	1.3525	0.4553	4.0176	No
Bloodpress	0.09	1	0.761938	0.8011	0.1909	3.3629	No

Source: The researcher output using NCSS software.

Mantel-Haenszel (M-H) statistics was illustrated in table (4.6.1), it showed that the following risk factors : PSA, age, marital status, diseases, alcohol, weight, intake of green vegetables and intake animal fat were significantly associated with prostate cancer incidence (the outcome variable) with; ($\chi^2 value$ =121.11, with p-value 0.000000), ($\chi^2 value$ =54.68, with p-value 0.000000), ($\chi^2 value$ =26.25, with p-value 0.000000), ($\chi^2 value$ =13.77, with p-value 0.000207), ($\chi^2 value$ =10.05, with p-value 0.001523), ($\chi^2 value$ =5.33, with p-value 0.021021), ($\chi^2 value$ =4.75, with p-value 0.029287) and ($\chi^2 value$ =4.05, with p-value 0.044048) respectively. Other variables that were not related to the outcome variable were: occupation, family history, cholesterol, blood pressure, intake of prostate medication and smoking are insignificant. In Table (4.6.1) all the risk factors were ranked by their importance. The state variable had been excluded because it consisted of 4 categories, and this test is concerned only with variables of two categories. The age variable was used as a confounding variable which was classified into two (categories) strata,

stratum1 represented age under 50 and stratum 2 represented age equal to or above 50. The highest MH odds ratio was 173.1 with confidence interval (69.1 - 433.4) and one degree of freedom for the variable PSA. This implies that men with abnormal PSA were more susceptible to the disease when the age was greater than or equal 50. So PSA was important risk factor for prostate cancer incidence.

Mantel-Haenszel procedure reported three sections for all variables under study to test independence between the disease status which categorized into two categories 0 represented controls and 1 represented cases (as we mentioned in the section of introduction) and the risk factor among strata of age variable, it was used as confounding variable with two strata; stratum1 represented men under the age of 50, and stratum 2 represented men equal or above 50. Next results illustrated these sections:

1- Occupation:

Table (4.23)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	81	4	10	1	2.0250
2	≥ 50	8	83	1	62	5.9759

Source: The researcher output using NCSS software.

A: occupation = 0(group1) , diagnostic = 0(without disease)

B: occupation = 0(group1) , diagnostic = 1(with disease)

C: occupation = 1(group2) , diagnostic = 0(without disease)

D: occupation = 1(group2) , diagnostic = 1(with disease)

Occupation variable had two categories; (0) represented group1 which pointed to males who had jobs depend on prolonged seating and (1) represented group2 which pointed to males with jobs depend on physical activities. Table (4.6.2) showed that each row of the report represents an individual 2-by-2 table and the definitions of the four letters (A, B, C, and D) were shown immediately below the table. Strata1 represented men under 50 of whom A (account 81) was the number of men without prostate cancer and they belonged to group1, also B (account 4) was the number of infected

men who belonged to group1. C (account 10) represented count of healthy men who their occupation pointed to group2. While D (account 1) clarified the count of men with cancer and occupation was classified into gorup2. Whilst Strata2 represented men equal to or above 50 of whom 8 men without disease and their occupation belonged to group1, also (83) was the number of infected men who belonged to group1. Only one man with no cancer and his occupation pointed to group2. While (62) clarified the count of men with prostate cancer and occupation was classified into gorup2.

Sample odds ratio is the odds ratio calculated for the 2-by-2 table listed on each row. In men less than 50, the incidence of prostate cancer increased in men who their occupation belonged to group2 (2.0250) times than those in group1. Also in men equal to or above 50, the incidence of the disease increased in men who their occupation was classified in group2 (6) times than those in group1. We noted that the jobs which belonged to group2 increased the incidence of prostate cancer in both strata.

Table (4.24)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.0782	2.3314	23.3437	0.4634	0.1146	0.0521
2	0.7267	4.9351	130.7197	0.0830	0.4091	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.3) presented results for an individual 2-by-2 table. The strata number provided the identity of particular 2-by-2 table, since the tables in this report were listed in the same order as those in the Strata count section report described previously. 1/2-Corrected Odds Ratio is defined when one or more cell counts are zero. In stratum1 the corrected odds ratio was 2.3314 and 4.9351 in stratum2, it differed from odds ratios in table (4.5.3). 95% confidence interval for the corrected odds ratio was (0.078-23.34). To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was 0.4634 and 0.0830 in sratum1 and stratum2 respectively. The both values were greater than 0.05, so the null hypothesis was accepted. We concluded that there was no clear effect of the nature of the work on prostate cancer incidence in both strata.

This table clarified that the proportion exposed in stratum2 (0.4091) was greater than proportion exposed in stratum1 (0.1146). Also the proportion

diseased in stratum2 (0.9416) was greater than the proportion diseased in stratum1 (0.0521). These proportions illustrated that the incidence of prostate cancer increased in men above 50 who had jobs which classified in group2.

Table (4.25)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. level
MH C.C.	0.7495	4.2533	24.1350	2.67	1	0.102160
MH	0.9909	4.2533	18.2556	3.79	1	0.051448
Woolf	0.7732	3.6388	17.1252	2.67	1	0.102169
Heterogeneity Test				0.47	1	0.495045

Source: The researcher output using NCSS software.

The MHCC row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction in table (4.6.4). The estimated odds ratio was 4.2533 with 95% C.I. (0.75 – 24.14). The Mantel-Haenszel χ^2 value was 2.67. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.102160) was greater than (0.05), so the null hypothesis cannot be rejected. This required that all odds ratios equal to one (OR in stratum1= OR in stratum2=1). MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 4.25 with 95% CI (0.99 – 18.26). The prob. level was greater than 0.05, so we accepted the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 3.64 with 95% C.L (0.77 – 17.13). The probability level was greater than 0.05, so we accepted the null hypothesis, hence the odds ratios were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.495045) was greater than 0.05, so we cannot reject the null hypothesis, therefore all odds ratios were equal.

2- Marital status:

Marital status variable had two categories; 0 represented unmarried men, and 1 represented married men. Table (4.6.5) clarified that each row of

the report represents an individual 2-by-2 table. The definitions of the four letters (A, B, C, and D) are shown immediately below the table.
In strata1 the unmarried men with prostate cancer were 5, while 35 men without cancer. Whilst there were not married men with the disease, also this table showed that 56 were married and without prostate cancer. In strata2 the unmarried men with prostate cancer were 141, while 5 men without. While married men with prostate cancer were 4 and 4 were healthy. Sample odds ratio is the odds ratio calculated for the 2-by-2 table listed on each row. The odds ratio in stratum1 (men less than 50) was not computable because this row included zero cell (cell D). In men equal to or above 50, the odds ratio was 0.0355, we concluded that married men suffered from prostate cancer more than unmarried men.

Table (4.26)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	35	5	56	0	Not exist
2	≥ 50	5	141	4	4	0.0355

Source: The researcher output using NCSS software.

A: marital status = 0(group1), diagnostic = 0(without disease)

B: marital status = 0(group1), diagnostic = 1(with disease)

C: marital status = 1(group2), diagnostic = 0(without disease)

D: marital status = 1(group2), diagnostic = 1(with disease)

Table (4.27)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.0120	0.0298	0.7818	0.0108	0.5833	0.0521
2	0.0049	0.0372	0.2345	0.0004	0.0519	0.9416

Source: The researcher output using NCSS software.

Each line on this report presents results for an individual 2-by-2 table. The strata number provided the identity of particular 2-by-2 table, since the

tables in this report were listed in the same order as those in the Strata count section report described previously. In table (4.6.6) the effective corrected odds ratio in stratum1 was 0.0298, with 95% C.I. (0.0120-0.7818). And the corrected odds ratio for stratum2 was 0.0372 with 95% C.I. (0.0049-0.2345). To test H_0 that the odds ratio was equal one, fisher exact test was used. The p-value of this test was (0.0108) and (0.0004) in stratum1and stratum2 respectively. Both p-values were less than (0.05), so the null hypothesis was rejected hence the ($OR \neq 1$) for both strata. We concluded that the marital status had effect on the prostate cancer incidence. The Proportion exposed (unmarried men) in stratum1 (0.5833) was greater than Proportion exposed in stratum2 (0.0519) in stratum2. Also the overall proportion of those in the table that were diseased in stratum2 (0.9416) was greater than proportion diseased (0.0521) in stratum1. These proportions showed that prostate cancer increased in married men above 50.

Table (4.28)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	0.0039	0.0197	0.1005	22.34	1	0.000002
MH	0.0044	0.0197	0.0886	26.25	1	0.000000
Woolf	0.0075	0.0346	0.1591	18.68	1	0.000015
Heterogeneity Test				0.01	1	0.938146

Source: The researcher output using NCSS software.

The MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 0.0197 with 95% C.I. (0.0039 – 0.1005), table (4.6.7). The Mantel-Haenszel χ^2 value was 22.34. It tested that the individuals stratum odds ratios are all equal to one versus the hypothesis that at least one odds ratio is different from unity. The probability level (0.000002) was less than (0.05), so the null hypothesis was rejected. This required that at least one odds ratio was not equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 0.0197 with 95% CI (0.0044 – 0.0886). The prob. level (0.000000) was less than 0.05, so we reject the null hypothesis that all odds ratios are equal to one, so at least one odds ratio differed from unity. The

estimated odds ratio of Woolf was 0.0346 with 95% C.L (0.0075 – 0.1591). The prob. Level (0.000015) was less than 0.05, so we reject the null hypothesis; hence at least one odds ratio not equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.938146) was greater than 0.05, so we accepted the null hypothesis, therefore all odds ratios were equal among strata.

3- Family history:

Table (4.29)

Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	74	4	17	1	1.0882
2	≥ 50	7	90	2	55	2.1389

Source: The researcher output using NCSS software.

A: famhist = 0(No), diagnostic = 0(without disease)

B: famhist = 0(No), diagnostic = 1(with disease)

C: famhist = 1(Yes), diagnostic = 0(without disease)

D: famhist = 1(Yes), diagnostic = 1(with disease)

Family history variable had two categories; 0 represented negative family history (family without prostate cancer), and 1 represented positive family history (infected relatives). Table (4.6.8) clarified that each row of the report represents an individual 2-by-2 table followed by the definitions of the four letters (A, B, C, and D). In strata1 the count of men with negative family history who infected with prostate cancer was 4, while 74 of men without cancer. The count of men without prostate cancer who had infected relatives (positive family history) was 17, whilst one man had healthy relatives. In a similar way the strata2 had 90 men suffered from prostate cancer with negative family history and 55 with positive family history. While 7 men without prostate cancer and they had healthy relatives, but two men with infected relatives. The Sample odds ratio is the odds ratio calculated for the 2-by-2 table listed on each row. The odds ratio was 1.2 and 2.1 in stratum1 and 2 respectively. So, when the age under 50 the odds of infected men equal to odds of healthy men. The prostate cancer incidence increased in

men over 50 who had infected relatives two times compared to those with healthy relatives.

Table (4.30)

Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.0434	1.2660	11.6383	1.0000	0.1875	0.0521
2	0.3864	1.9726	15.5013	0.4861	0.3701	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.9) presents results for an individual 2-by-2 table. The corrected odds ratio was 1.3 and 1.97 in strata 1 and 2 respectively, and 95% confidence interval for the corrected odds ratio was conducted. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was greater than (0.05) in each stratum, so the null hypothesis cannot be rejected. We concluded that $OR = 1$ for each stratum. Thus, family history had no effect on prostate cancer incidence. The proportion exposed in stratum1 was (0.1875) less than (0.3701) stratum2. And proportion diseased in men less than 50 was (0.0521) less than proportion diseased (0.9416) in men above 50. We conclude that the presence of infected relatives with prostate cancer increased the appearance of the disease in men above 50.

Table (4.31)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	0.2340	1.7424	12.9758	0.29	1	0.587783
MH	0.4809	1.7424	6.3139	0.71	1	0.397929
Woolf	0.4603	1.7033	6.3025	0.64	1	0.424999
Heterogeneity Test				0.23	1	0.632314

Source: The researcher output using NCSS software.

Table (4.6.10) showed that the MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 1.7424 with 95% C.I. (0.23 – 12.98). The Mantel-Haenszel χ^2 value was 0.29. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.587783) was greater than 0.05, so the null hypothesis cannot be rejected. This required that all odds ratios equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 1.7424 with 95% CI (0.48 – 6.31). The prob. level (0.397929) was greater than 0.05, so we accepted the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 1.7033 with 95% C.L (0.46 – 6.30). The prob. Level (0.424999) was greater than 0.05, so we didn't reject the null hypothesis; hence all odds ratio were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.632314) was greater than 0.05, so we accepted the null hypothesis, therefore all odds ratios were equal.

4- Animal fat and red meat:

Table (4.32)

Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	35	0	56	5	Not exist
2	≥ 50	3	26	6	119	2.2885

Source: The researcher output using NCSS software.

A: redmeat = 0(No), diagnostic = 0(without disease)

B: redmeat = 0(No), diagnostic = 1(with disease)

C: redmeat = 1(Yes), diagnostic = 0(without disease)

D: redmeat = 1(Yes), diagnostic = 1(with disease)

This study also studied intake of animal fat and red meat regularly; "0" represented balanced diet and "1" represented diet containing high animal

fats. Table (4.6.11) clarified that each row of the report represents an individual 2-by-2 table followed by the definitions of the four letters (A, B, C, and D). In stratum1 the count of men without prostate cancer and they followed a balanced diet was 35 and 56 did not follow. Of all those suffering from prostate cancer, we found that five men had a diet without animal fat, while none of them depended on animal fats. In stratum2 the number of men without prostate cancer and they followed a balanced diet was 3 and 6 did not follow. Of all those suffering from prostate cancer, we found that 26 men had a diet without animal fat, while 119 depended on animal fat. The odds ratio for sratum2 was (2.3).

Table (4.33)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.4993	13.1600	32.4719	0.1549	0.6354	0.0521
2	0.4202	2.3623	11.3068	0.3716	0.8117	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.12) presents results for an individual 2-by-2 table. The corrected odds ratio was nearly 1 and 2.4 in strata 1 and 2 respectively, and 95% confidence interval for the corrected odds ratio was conducted. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was greater than (0.05) in each stratum, so the null hypothesis cannot be rejected. We concluded that $OR = 1$ for each stratum. Thus, eating of the animal fats and red meat regularly had no effect on prostate cancer incidence. The corrected odds ratio for stratum1 was 13.2, it is more reliable than odds ratio in table (4.5.9). The proportion exposed and diseased in stratum1 were (0.6354) and (0.0521) respectively. Also the proportion exposed and diseased in stratum2 were (0.8117) and (0.9416) respectively. We concluded that the proportion exposed and proportion diseased in men above or equal 50 were greater than those in men less than50. So eating of red meats and animal fat increased the appearance of prostate cancer in men above 50.

Table (4.34)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	0.7999	4.0880	20.8913	2.86	1	0.090690
MH	1.0382	4.0880	16.0963	4.05	1	0.044048
Woolf	0.7145	2.7951	10.9343	2.18	1	0.139695
Heterogeneity Test				0.64	1	0.423815

Source: The researcher output using NCSS software.

In table (4.6.13), the MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 4.0880 with 95% C.I. (0.80 – 20.89). The Mantel-Haenszel χ^2 value was 2.86. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.090690) was greater than 0.05, so the null hypothesis cannot be rejected. This required that all odds ratios equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 4.0880with 95% C.I. (1.038 – 16.096). The prob. level (0.044048) was less than 0.05, so we reject the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 2.7951 with 95% C.L (0.715 – 10.934). The prob. Level (0.139695) was greater than 0.05, so we didn't reject the null hypothesis; hence all odds ratio were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.423815) was greater than 0.05, so we accepted the null hypothesis, therefore all odds ratios were equal.

5- Green vegetables and fruits:

Table (4.35)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	68	4	23	1	0.7391
2	≥ 50	8	60	1	85	11.3333

Source: The researcher output using NCSS software.

A: greenveg = 0(No) , diagnostic = 0(without disease)

B: greenveg = 0(No) , diagnostic = 1(with disease)

C: greenveg = 1(Yes) , diagnostic = 0(without disease)

D: greenveg = 1(Yes) , diagnostic = 1(with disease)

In this table we discussed the intake of green vegetables and fruits regularly; "0" indicated that the diet contains green fruits and vegetables and "1" indicated that the diet did not contain these nutrients. Table (4.6.14) showed that each row of the report represents an individual 2-by-2 table. Regarding to stratum1 of all those without prostate cancer we noted that 68 men did not regularly enter fruits and vegetables into their diet, while 23 did. In this stratum the number of men who suffered from prostate cancer was five. Of whom 4 men did not preferred fruits and vegetables while one man had good diet. Regarding to stratum2 of all those without prostate cancer we noted that 8 men did not regularly enter fruits and vegetables into their diet, while one man did. In this stratum the number of men who suffered from prostate cancer was 145, of whom 60 men did not preferred fruits and vegetables while 85 men had good diet. The odds ratio for stratum1 and 2 were 0.74 and 11.3 respectively.

Table (4.36)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.0299	0.8634	7.6999	1.0000	0.7500	0.0521
2	1.3778	9.3386	247.9934	0.0108	0.4416	0.9416

Source: The researcher output using NCSS software.

The table (4.6.15) illustrates that the corrected odds ratio was nearly 1 and 2.4 in strata 1 and 2 respectively, and 95% confidence interval for the corrected odds ratio was conducted. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was greater than (0.05) in stratum1, so the null hypothesis cannot be rejected. We

concluded that $OR = 1$, so the eating of fruits and green vegetables had no effect on prostate cancer incidence. but in stratum2 the p-value was less than 0.05, so the null hypothesis was rejected and we concluded that these nutrients had effect on the disease in men greater than50. The proportions exposed and diseased in stratum1 were (0.7500) and (0.0521) respectively. Also the proportions exposed and diseased in stratum2 were (0.4416) and (0.9416) respectively. The proportion exposed in men less than 50 was greater than men above 50. And the proportion diseased in men above 50 was greater than those under 50 years. From these proportions we conclude that eating of fruits and green vegetables reduced the risk of prostate cancer.

Table (4.37)

Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	0.9525	3.8013	15.1710	3.58	1	0.058626
MH	1.1440	3.8013	12.6305	4.75	1	0.029287
Woolf	0.6798	3.1536	14.6297	2.15	1	0.142377
Heterogeneity Test				3.03	1	0.081853

Source: The researcher output using NCSS software.

The MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 3.8013 with 95% C.I. (0.95 – 15.17), table (4.6.16). The Mantel-Haenszel χ^2 value was 3.58. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.058626) was greater than 0.05, so the null hypothesis cannot be rejected. This required that all odds ratios equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 3.8013 with 95% C.I. (1.14 – 12.63). The prob. level (0.029287) was less than 0.05, so we reject the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 3.1536 with 95% C.L (0.68 – 14.63). The prob. Level (0.142377) was greater than 0.05, so we didn't reject the null hypothesis; hence all odds

ratio were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.081853) was greater than 0.05, so we accepted the null hypothesis, therefore all odds ratios were equal.

6- Overweight:

Table (4.38)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	74	1	17	4	17.4118
2	≥ 50	7	98	2	47	1.6786

Source: The researcher output using NCSS software.

A: overweight = 0(No) , diagnostic = 0(without disease)
B: overweight = 0(No) , diagnostic = 1(with disease)
C: overweight = 1(Yes) , diagnostic = 0(without disease)
D: overweight = 1(Yes) , diagnostic = 1(with disease)

The overweight variable categorized into two categories; 0 for normal weight and 1 for overweight. Table (4.6.17) showed that each row of the report represents an individual 2-by-2 table. Regarding to stratum1 of all those without prostate cancer we noted that 74 men had normal weight, while 17 men suffered from overweight. In this stratum the number of men who suffered from prostate cancer was five, of whom 4 men were suffered from obesity while one man had normal weight. Regarding to stratum2 of all those without prostate cancer, we noted that 7 men did not suffer from overweight, while two men did. In this stratum the number of men who suffered from prostate cancer was 145, of whom 98 men had normal weight and 47 were suffered from obesity. The odds ratio for stratum1 and 2 were 17.4118 and 1.6786 respectively.

Table (4.39)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	1.6408	14.6348	437.2872	0.0077	0.2188	0.0521
2	0.3018	1.5496	12.2013	0.7197	0.3182	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.18) presents results for an individual 2-by-2 table. The corrected odds ratios were 14.63 with 95% CI: (1.64 - 437.29) and 1.55 with 95% CI: (0.301 – 12.20) in strata 1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was less than (0.05) in stratum1, so we reject the null hypothesis. We concluded that $OR \neq 1$ for stratum1. Because the probability level of fisher exact test was greater than (0.05) in stratum2, we cannot reject the null hypothesis. We concluded that $OR = 1$ for stratum2. Thus suffering from obesity in men less than 50 had effect on prostate cancer incidence. In stratum1 (men under 50) proportion exposed and proportion diseased were 0.2188 and 0.0521 respectively. But in stratum2 (men above than 50) the Proportion exposed and proportion diseased was 0.3182 and 0.9416 respectively. The overall proportions of exposed and diseased men in stratum2 were greater than those in stratum1. Hence the obesity had a clear role in appearance of prostate cancer in men above 50.

Table (4.40)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	1.0270	3.6003	12.6215	4.01	1	0.045333
MH	1.2129	3.6003	10.6870	5.33	1	0.021021
Woolf	0.9982	3.6986	13.7050	3.83	1	0.050322
Heterogeneity Test				2.74	1	0.097835

Source: The researcher output using NCSS software.

In table (4.6.19) the MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 3.6003 with 95% C.I. (1.03 – 12.60). The Mantel-Haenszel χ^2 value was 4.01. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.045333) was less than 0.05, so the null hypothesis was rejected. This required that at least there was one odd ratio differed from one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 3.6003 with 95% C.I. (1.21 – 10.69). The prob. level (0.021021) was less than 0.05, so we reject the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 3.6986 with 95% C.L (0.998 – 13.705). The prob. Level (0.050322) was greater than 0.05, so we didn't reject the null hypothesis; hence all odds ratio were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.097835) was greater than 0.05, so we accepted the null hypothesis, therefore all odds ratios were equal.

7- High cholesterol:

Table (4.41)

Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	89	4	2	1	11.1250
2	≥ 50	8	105	1	40	3.0476

Source: The researcher output using NCSS software.

A: cholest = 0(No) , diagnostic = 0(without disease)
B: cholest = 0(No) , diagnostic = 1(with disease)
C: cholest = 1(Yes) , diagnostic = 0(without disease)
D: cholest = 1(Yes) , diagnostic = 1(with disease)

The high cholesterol variable was categorized into two categories; 0 for normal cholesterol and 1 for high cholesterol. Table (4.6.20) illustrated that each row of the report represents an individual 2-by-2 table. The number of men without prostate cancer in stratum1 was 91of whom 89 had normal cholesterol and two were suffered from

high cholesterol. Five men diagnosed with prostate cancer, four of whom had normal cholesterol while one had not. The total number of men who did not suffer from prostate cancer in stratum2 was 9 of whom 8 men with normal cholesterol and one without. Whilst the total number of men who suffered from this cancer 145 of whom 105 had normal cholesterol and 40 had not. The odds ratios for both stratum1 and 2 were 11.125 and 3.0476 respectively.

Table (4.42)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.3229	11.6667	236.3015	0.1497	0.0313	0.0521
2	0.3661	2.5240	67.0376	0.4461	0.2662	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.21) presents results for an individual 2-by-2 table. The corrected odds ratios were 11.67 with 95% CI: (0.32 – 236.30) and 2.52 with 95% CI: (0.37 – 67.04) in strata 1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was greater than (0.05) in both strata, so we cannot reject the null hypothesis. We concluded that $OR = 1$ in stratum1 and stratum2. In men under50 the proportion exposed and proportion diseased were 0.0313 and 0.0521 respectively. But in men above than 50 the Proportion exposed and proportion diseased was 0.2662 and 0.9416 respectively. The overall proportions of exposed and diseased men in stratum2 were greater than those in stratum1. These proportions showed us that high cholesterol had effect on men over 50 years, where the risk of developing prostate cancer increases.

Table (4.43)

Mantel-Haenszel Statistics Section

Method	Lower 95%CL	Estimated odds ratio	Upper 95% CL	χ^2 value	df	Prob. Level
MH C.C.	0.4936	3.9273	31.2465	1.67	1	0.196087
MH	0.7843	3.9273	19.6670	2.77	1	0.096053
Woolf	0.9896	5.0957	26.2401	3.79	1	0.051484
Heterogeneity Test				0.57	1	0.448663

Source: The researcher output using NCSS software.

The MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 3.9273 with 95% C.I. (0.49 – 31.25), table (4.6.22). The Mantel-Haenszel χ^2 value was 1.67. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.196087) was greater than 0.05, so the null hypothesis cannot be rejected. This required that all odds ratios were equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 3.9273 with 95% C.I. (0.78 – 19.67). The prob. level (0.096053) was greater than 0.05, so we accepted the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 5.0957 with 95% C.L (0.99 – 26.24). The prob. Level (0.051484) was greater than 0.05, so we didn't reject the null hypothesis; hence all odds ratio were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.448663) was greater than 0.05, so we accepted the null hypothesis, therefore all odds ratios were equal.

8- High blood pressure:

Table (4.44)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	89	5	2	0	Not exist
2	≥ 50	6	102	3	43	0.8431

Source: The researcher output using NCSS software.

A: bloodpres = 0(No) , diagnostic = 0(without disease)
B: bloodpres = 0(No) , diagnostic = 1(with disease)
C: bloodpres = 1(Yes), diagnostic = 0(without disease)
D: bloodpres =1(Yes), diagnostic = 1(with disease)

The high blood pressure variable was categorized into two categories; 0 for normal pressure and 1 for high pressure. Table (4.6.23) illustrated that each row of the report represents an individual 2-by-2 table. The number of men without prostate cancer in stratum1 was 91of whom 89 had normal blood pressure and two were suffered from high blood pressure. Five men diagnosed with prostate cancer, and they had normal blood pressure. The total number of men who did not suffer from prostate cancer in stratum2 was 9 of whom 6 men with normal blood pressure and 3 without. Whilst the total number of men who suffered from this cancer 145 of whom 102 had normal blood pressure and 43 had not. The odds ratio for stratum2 was 0.8431. The odds ratio in stratum1 cannot be calculated because cell D was zero.

Table (4.45)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.9686	1.8889	98.9495	1.0000	0.0208	0.0521
2	0.1758	0.8134	4.4894	1.0000	0.2987	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.24) presents results for an individual 2-by-2 table. The corrected odds ratios were 1.89 with 95% CI: (0.97 – 98.95) and 0.81 with 95% CI: (0.18 – 4.45) in strata 1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was greater than (0.05) in both strata, so we cannot reject the null hypothesis. We concluded that $OR = 1$ in stratum1 and stratum2. The proportions exposed were 0.0208 and 0.2987 in stratum1 and stratum2 respectively, so the exposed proportion (proportion of men who suffered from high blood pressure) in stratum2 was greater than its counterpart in stratum1. The proportions diseased were 0.0521 and 0.9416 in stratum1 and stratum2 respectively, so the proportion diseased in stratum2 was greater than stratum1. We noted that the high blood pressure had a role in appearance of prostate cancer in men above 50.

Table (4.46)
Mantel-Haenszel Statistics Section

Method	Lower95% C.L.	Estimated odds ratio	Upper95% C.L.	χ^2value	df	Prob. Level
MH C.C.	0.0007	0.8011	960.8424	0.00	1	0.951123
MH	0.1909	0.8011	3.3629	0.09	1	0.761938
Woolf	0.2362	0.9161	3.5528	0.02	1	0.899188
Heterogeneity Test				0.13	1	0.723014

Source: The researcher output using NCSS software.

The MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. Table (4.6.25) showed the estimated odds ratio was 0.8011 with 95% C.I. (0.001 – 960.842). The Mantel-Haenszel χ^2 value was 0.00. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.951123) was greater than 0.05, so the null hypothesis cannot be rejected. This required that all odds ratios were equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without

continuity correction. The estimated odds ratio for MH test was 0.8011 with 95% C.I. (0.191 – 3.363). The prob. level (0.761938) was greater than 0.05, so we accepted the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 0.9161 with 95% C.L (0.236 – 3.553). The prob. Level (0.899188) was greater than 0.05, so we didn't reject the null hypothesis; hence all odds ratio were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.723014) was greater than 0.05, so we accepted the null hypothesis, therefore all odds ratios were equal.

9- Intake of prostate medications:

Table (4.47)

Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	2	2	89	3	0.0337
2	≥ 50	3	50	6	95	0.9500

Source: The researcher output using NCSS software.

A: prostatmed = 0(No), diagnostic = 0(without disease)
B: prostatmed = 0(No), diagnostic = 1(with disease)
C: prostatmed = 1(Yes), diagnostic = 0(without disease)
D: prostatmed = 1(Yes), diagnostic = 1(with disease)

The Intake of prostate medications was categorized into two categories; (0 for No) and (1 for Yes). Table (4.6.26) showed that each row of the report represents an individual 2-by-2 table. The number of men without prostate cancer in stratum1 was 91of whom 89 took medications to treat prostate diseases before cancer appeared and two did not. Five men diagnosed with prostate cancer of whom 3 men took prostate medication and two didn't. The total number of men who did not suffer from prostate cancer in stratum2 was 9 of whom 6 men took prostate medication and 3did not take. Whilst the total number of men who suffered from this cancer was 145 of whom 50 men did not have prostate problems before cancer appeared and 95 suffered from prostatitis or prostate enlargement and took its

medications. The odds ratios for stratum1 and stratum2 were 0.0337 and 0.9500 respectively.

Table (4.48)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.0019	0.0364	0.4943	0.0126	0.0417	0.0521
2	0.1792	0.9857	4.5305	1.0000	0.3442	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.27) presents results for an individual 2-by-2 table. The corrected odds ratios were 0.0364 with 95% CI: (0.0019 – 0.4943) and 0.9857 with 95% CI: (0.1792 – 4.5305) in strata 1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was less than (0.05) in stratum1, so we reject the null hypothesis. We concluded that $OR \neq 1$ in stratum1. In stratum2 the probability level was greater than (0.05), so we don't reject the null hypothesis. We concluded that $OR = 1$. In stratum1 proportion exposed and proportion diseased were 0.0417 and 0.0521 respectively. But in men greater than 50 the Proportion exposed and proportion diseased was 0.3442 and 0.9416 respectively. In stratum1 the proportion exposed and proportion diseased were less than those in stratum2. These proportions showed that men (above 50) who did not take prostate medication more susceptible to prostate cancer incidence.

In table (4.6.28) the MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 0.5032 with 95% C.I. (0.123 – 2.061).

Table (4.49)
Mantel-Haenszel Statistics Section

Method	Lower 95%CL	Estimated OR	Upper 95% CL	χ^2 value	df	Prob. Level
MH C.C.	0.1229	0.5032	2.0607	0.91	1	0.339670
MH	0.1784	0.5032	1.4189	1.69	1	0.194113
Woolf	0.1103	0.3694	1.2373	2.61	1	0.106373
Heterogeneity Test				5.94	1	0.014761

Source: The researcher output using NCSS software.

The Mantel-Haenszel χ^2 value was 0.91, it tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.339670) was greater than 0.05, so the null hypothesis cannot be rejected. This required that all odds ratios were equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 0.5032 with 95% C.I. (0.178– 1.419). The prob. level (0.194113) was greater than 0.05, so we accepted the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 0.3694 with 95% C.L (0.110 – 1.237). The prob. Level (0.106373) was greater than 0.05, so we didn't reject the null hypothesis; hence all odds ratio were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.014761) was less than 0.05, so we rejected the null hypothesis, thus at least one odd ratio doddered from the others.

10- Alcohol:

Table (4.50)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	85	2	6	3	21.2500
2	≥ 50	8	88	1	57	5.1818

Source: The researcher output using NCSS software.

A: alcohol = 0(No) , diagnostic = 0(without disease)

B: alcohol = 0(No) , diagnostic = 1(with disease)
C: alcohol = 1(Yes), diagnostic = 0(without disease)
D: alcohol = 1(Yes), diagnostic = 1(with disease)

Alcohol consumption variable was categorized into two categories; (0 for No) and (1 for Yes). Table (4.6.29) showed that each row of the report represents an individual 2-by-2 table. The number of men without prostate cancer in stratum1 was 91of whom 85 men took alcohol before the onset of cancer and six did not. Five men diagnosed with prostate cancer of whom 3 men took alcohol and two didn't. The total number of men who did not suffer from prostate cancer in stratum2 was 9 of whom one man took alcohol and 8 men did not take. Whilst the total number of men who suffered from this cancer was 145 of whom 88 men did not take alcohol problems before cancer appeared and 57 did. The odds ratios for stratum1 and stratum2 were 21.2500 and 5.1818 respectively.

Table (4.51)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected OR	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	2.2355	19.7022	238.9165	0.0053	0.0938	0.0521
2	0.6293	4.2816	113.4215	0.1544	0.3766	0.9416

Source: The researcher output using NCSS software.

Each in table (4.6.30) presents results for an individual 2-by-2 table. The corrected odds ratios were 19.7 with 95% CI: (2.24 – 238.92) and 4.28 with 95% CI: (0.63 – 113.42) in strata 1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was less than (0.05) in stratum1, so we reject the null hypothesis. We concluded that $OR \neq 1$ in stratum1. In stratum2 the probability level was greater than (0.05), so we don't reject the null hypothesis. We concluded that $OR = 1$. In stratum1 proportion exposed and proportion diseased were 0.0938 and 0.0521 respectively. But in men greater than 50 the Proportion exposed and proportion diseased was 0.3766 and 0.9416

respectively. It was clear from these proportions that alcohol has an impact on appearance of prostate cancer in men above 50.

Table (4.52)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	1.9174	8.0659	33.9301	8.11	1	0.004399
MH	2.2188	8.0659	29.3216	10.05	1	0.001523
Woolf	2.6065	10.9909	46.3454	10.66	1	0.001096
Heterogeneity Test				0.92	1	0.337588

Source: The researcher output using NCSS software.

The MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 8.0659 with 95% C.I. (1.92 – 33.93), table (4.6.31). The Mantel-Haenszel chi-square value was 8.11. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.004399) was less than 0.05, so we rejected the null hypothesis. This required that at least one odd ratios not equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 8.0659 with 95% C.I. (2.22– 29.32). The prob. level (0.001523) was less than 0.05, so we rejected the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 10.9909 with 95% C.L (2.61 – 46.35). The prob. Level (0.001096) was less than 0.05, so we rejected the null hypothesis; hence one of the odds ratios not equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.337588) was greater than 0.05, so we accepted the null hypothesis, that all odds ratios were equal.

11- Smoking:

Table (4.53)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	39	2	52	3	1.1250
2	≥ 50	5	66	4	79	1.4962

Source: The researcher output using NCSS software.

A: smoking = 0(No) , diagnostic = 0(without disease)
B: smoking = 0(No) , diagnostic = 1(with disease)
C: smoking = 1(Yes), diagnostic = 0(without disease)
D: smoking = 1(Yes), diagnostic = 1(with disease)

The smoking variable was categorized into two categories; (0 for No) and (1 for Yes). Table (4.6.32) showed that each row of the report represents an individual 2-by-2 table. The number of men without prostate cancer in stratum1 was 91of whom 52 were smokers and 39 non smokers. Five men diagnosed with prostate cancer of whom 3 men were smokers and two were non smokers. The total number of men who did not suffer from prostate cancer in stratum2 was 9 of whom 5 men were non smokers and 4 were smokers. Whilst the total number of men who suffered from this cancer was 145 of whom 66 men were non smokers and 79 were smokers. The odds ratios for stratum1 and stratum2 were 1.1250 and 1.4962 respectively.

Table (4.54)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	0.1429	1.0851	10.2021	1.0000	0.5792	0.0521
2	0.3309	1.4777	6.9740	0.7334	0.5389	0.9416

Source: The researcher output using NCSS software.

Each line table (4.6.33) presents results for an individual 2-by-2 table. The corrected odds ratios were 1.09 with 95% CI: (0.14 –

10.20) and 1.48 with 95% CI: (0.33 – 6.97) in strata 1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was greater than (0.05) in both strata, so we don't reject the null hypothesis. We concluded that $OR = 1$ in both strata. In stratum1 proportion exposed and proportion diseased were 0.5792 and 0.0521 respectively. But in stratum2 the Proportion exposed and proportion diseased was 0.5389 and 0.9416 respectively. Hence the exposed proportion in stratum1 was greater than its counterpart in stratum2, but the proportion diseased in stratum1 was less than stratum2. We conclude that there was no clear effect for smoking in the appearance of prostate cancer.

Table (4.55)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	0.1486	1.3525	12.3131	0.07	1	0.788758
MH	0.4553	1.3525	4.0176	0.30	1	0.586767
Woolf	0.4548	1.3532	4.0259	0.30	1	0.586636
Heterogeneity Test				0.06	1	0.806552

Source: The researcher output using NCSS software.

The MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 1.3525 with 95% C.I. (0.15 – 12.31), table (4.6.34). The Mantel-Haenszel chi-square value was 0.07. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.788758) was greater than 0.05, so we accepted the null hypothesis that all odds ratios were equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 1.3525 with 95% C.I. (0.46– 4.02). The prob. level (0.586767) was greater than 0.05, so we accepted the null

hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 1.3532 with 95% C.L (0.45 – 4.03). The prob. Level (0.586636) was greater than 0.05, so the null was accepted; hence all the odds ratios were equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.806552) was greater than 0.05, so we accepted the null hypothesis, that all odds ratios were equal.

12- Developing one or more of these diseases: "syphilis, gonorrhea, chronic prostatitis and prostate enlargement":

Table (4.56)

Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	88	2	3	3	44.0000
2	≥ 50	6	48	3	97	4.0417

Source: The researcher output using NCSS software.

A: diseases = 0(No), diagnostic = 0(without disease)
B: diseases = 0(No), diagnostic = 1(with disease)
C: diseases = 1(Yes), diagnostic = 0(without disease)
D: diseases = 1(Yes), diagnostic = 1(with disease)

The Developing one or more of these diseases: "syphilis, gonorrhea, chronic prostatitis and prostate enlargement" was categorized into two categories; (0 for No) and (1 for Yes). Table (4.6.35) showed that each row of the report represents an individual 2-by-2 table. The number of men without prostate cancer in stratum1 was 91of whom 88 men suffered from one or more of the above disease before cancer appeared and three men didn't suffer. Five men diagnosed with prostate cancer of whom two men didn't suffer from any of the above diseases and three men did. The total number of men without prostate cancer in stratum2 was 9 of whom 6 men didn't suffer from one of the above diseases and three men did. Whilst the total number of men who suffered from this cancer was 145 of whom 48 men did not suffer from these diseases before cancer appeared and

97 did. The odds ratios for stratum1 and stratum2 were 44.00 and 4.041 respectively.

Table (4.57)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected odds ratio	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	3.8172	39.2222	683.9698	0.0013	0.0625	0.0521
2	0.8472	3.8760	21.4724	0.0668	0.6494	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.36) represents results for an individual 2-by-2 table. The corrected odds ratios were 39.22 with 95% CI: (3.82 – 683.97) and 3.88 with 95% CI: (0.85 – 21.47) in strata 1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was less than (0.05) in stratum1, so we reject the null hypothesis. We concluded that $OR \neq 1$ in stratum1. In stratum2 the probability level was greater than (0.05), so we don't reject the null hypothesis. We concluded that $OR = 1$. The proportion exposed for stratum1 was 0.0625 and for stratum2 was 0.6494, so overall proportion of men who developed from one or more of the above diseases in stratum2 was greater than that proportion in stratum1. In a similar manner the proportion diseased was 0.0521 and 0.9416 in stratum1 and stratum2 respectively, so the proportion diseased in stratum2 was greater than stratum1. We conclude that developing one or more of these diseases; "syphilis, gonorrhea, chronic prostatitis and prostate enlargement" had a role in the appearance of prostate cancer in men above 50.

Table (4.58)
Mantel-Haenszel Statistics Section

Method	Lower 95% C.L.	Estimated odds ratio	Upper 95% C.L.	Chi-square value	df	Prob. Level
MH C.C.	2.1983	6.5452	19.4875	11.39	1	0.000738
MH	2.4261	6.5452	17.6574	13.77	1	0.000207
Woolf	2.5928	8.4876	27.7846	12.49	1	0.000408
Heterogeneity Test				3.34	1	0.067815

Source: The researcher output using NCSS software.

In table (4.6.37) the MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 6.5452 with 95% C.I. (2.20 – 19.49). The Mantel-Haenszel chi-square value was 11.39. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.000738) was less than 0.05, so we rejected the null hypothesis. This required at least one odds ratio were not equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 6.5452 with 95% C.I. (2.43– 17.66). The prob. level (0.000207) was less than 0.05, so we rejected the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 8.4876 with 95% C.L (2.60 – 27.78). The prob. Level (0.000408) was less than 0.05, so the null was rejected; hence at least one odds ratio not equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.067815) was greater than 0.05, so we accepted the null hypothesis, that all odds ratios were equal.

13- Prostate specific antigen (PSA):

Table (4.59)
Strata Count Section

Strata	Age	A	B	C	D	Sample odds ratio
1	< 50	88	0	3	5	Not exist
2	≥ 50	6	3	3	142	94.6667

Source: The researcher output using NCSS software.

A: PSA = 0(No), diagnostic = 0(without disease)
B: PSA = 0(No), diagnostic = 1(with disease)
C: PSA = 1(Yes), diagnostic = 0(without disease)
D: PSA = 1(Yes), diagnostic = 1(with disease)

The prostate specific antigen variable was categorized into two categories; 0 for normal PSA and 1 for abnormal PSA. Table (4.6.38) illustrated that each row of the report represents an individual 2-by-2 table. The number of men without prostate cancer in stratum1 was 91of whom 88 had abnormal PSA and three men had normal PSA. All five men diagnosed with prostate cancer suffered from abnormal PSA. The total number of men who did not suffer from prostate cancer in stratum2 was 9 of whom 6 men with normal PSA and 3 without. Whilst the total number of men who suffered from this cancer 145 of whom 3 had normal PSA and 142 had not. The odds ratio for stratum2 was 94.6667. The odds ratio in stratum1 cannot be calculated because cell B was zero.

Table (4.60)
Strata Detail Section

Strata	Lower 95% CL	1/2-Corrected OR	Upper 95% CL	Exact Test	Proportion Exposed	Proportion diseased
1	16.0951	570.2308	1419.2616	0.0000	0.0833	0.0521
2	12.4127	84.1716	942.9183	0.0000	0.9416	0.9416

Source: The researcher output using NCSS software.

Each line in table (4.6.39) represents results for an individual 2-by-2 table. The corrected odds ratios were 570.23 and 84.17 in strata

1 and 2 respectively. To test H_0 that the odds ratio is equal one, fisher exact test was used. The probability level of fisher exact test was less than (0.05) in both strata, so we reject the null hypothesis. We concluded that $OR \neq 1$ in stratum1 and stratum2. The proportion exposed for stratum1 was 0.0833 and for stratum2 was 0.9416, so in stratum2 the overall proportion of men who suffered from abnormal PSA was greater than that proportion exposed in stratum1. In a similar manner the proportion diseased was 0.0521 and 0.9416 in stratum1 and stratum2 respectively, so the proportion diseased in stratum2 was greater than stratum1. We conclude that abnormal PSA had effect on the incidence of prostate cancer in men above 50.

Table (4.61)
Mantel-Haenszel Statistics Section

Method	Lower 95% CL	Estimated odds ratio	Upper 95% CL	Chi-square value	df	Prob. Level
MH C.C.	65.8855	173.0926	454.7443	109.37	1	0.000000
MH	69.1265	173.0926	433.4235	121.11	1	0.000000
Woolf	24.1151	125.5160	653.2940	32.97	1	0.000000
Heterogeneity Test				0.60	1	0.437543

Source: The researcher output using NCSS software.

The MH.C.C row presented the Mantel-Haenszel confidence limits and hypothesis test with continuity correction. The estimated odds ratio was 173.0926 with 95% C.I. (65.89 – 454.74), table (4.6.40). The Mantel-Haenszel chi-square value was 109.37. It tested the null hypothesis that the individual stratum odds ratios are all equal to one versus the alternative hypothesis that at least one odds ratio is different from unity. The probability level (0.000000) was less than 0.05, so we rejected the null hypothesis. This required at least one odds ratio were not equal to one. MH row presented the Mantel-Haenszel confidence limits and hypothesis test without continuity correction. The estimated odds ratio for MH test was 173.0926 with 95% C.I. (69.13 – 433.42). The prob. level (0.000000) was less than

0.05, so we rejected the null hypothesis that all odds ratios are equal to one. The estimated odds ratio of Woolf was 125.5160 with 95% C.L (24.12 – 653.29). The prob. Level (0.000000) was less than 0.05, so the null was rejected; hence at least one odds ratio not equal to one. Heterogeneity row presented a hypothesis test developed by Woolf. The probability level (0.437543) was greater than 0.05, so we accepted the null hypothesis, that all odds ratios were equal.

(4.7) Discussion of the Results:

There was an agreement between the results of the chi-square test and Mantel-Haenszel test. The agreed risk factors were: PSA, age, diseases, alcohol, weight, animal fats, marital status and intake of green vegetables.

Abnormal PSA increases the risk of the disease with very large odds ratio (173.1) and χ^2- values 121.11 and 210.0101 form Table (4.6.1) and Table (4.5.1), respectively. So, PSA was most important variable in this study. Other studies had reached similar results; Ernesto, E.P. et al. (2016) conducted an analytical study of 218 Japanese patients. They had first developed a theoretical framework to study PSA dynamics for BPH and prostate cancer patients. This analytical study then was applied to obtain monograms for a better understanding of the relationship among PSA and tumor volume in Japanese men with proven BPH or proven prostate cancer. This novel approach which does not neglect PSA contribution due to BPH may provide new information useful for a better diagnostic and prognosis of prostatic diseases or localized prostate cancer. They provided a relationship among PSA, age, and tumor volume. Another study provided by Swanson, Kristin R. et al. (2001) developed a mathematical model for the dynamics of serum levels of PSA as a function of the tumor volume. Their model results show good agreement with experimental observations and provide an explanation for the existence of significant prostatic tumor mass despite a low-serum PSA. This result can be very useful in enhancing the use of serum PSA levels as a marker for cancer growth.

Also age was an important risk factor for prostate cancer incidence, with χ^2- values 191.2548 and 54.68 by using chi-square test and Mantel-

Haenszel test respectively. Men above the age of 50 years were exposed to the disease 56.2 times than those who were younger, Table (4.6.1). There were some studies that confirm validity of this study; Carter, H.B. et al. (1990) showed that 50% of men between 70 and 80 years of age showed histological evidence of malignancy. At that time risk of 42% for developing histological evidence of prostate cancer in 50-year-old men had been calculated. In men at this age, however, the risk of developing clinically significant disease was only 9.5%, and the risk of dying from prostate cancer was only 2.9%.

Marital status had effect to the disease with (p-value< 0.05) in both tables. Table (4.6.1) illustrated that married men were more susceptible than unmarried. However Tyson M.D. et al. (2013) found different results. They used Multivariate Cox regression techniques to study the relationship between marital status and prostate cancer and overall mortality. They concluded that marital status was an independent predictor of prostate cancer-specific mortality and overall mortality in men with prostate cancer. Unmarried men have a higher risk of prostate cancer-specific mortality compared to married men of similar age, race, stage, and tumor grade.

The variable of developing one or more of these diseases: "syphilis, gonorrhea, chronic prostatitis and prostate enlargement" was significantly related to the outcome variable. The values of chi-square test and Mantel-Haenszel test were 87.95 (p-value 0.0000) and 13.77 (p-value 0.0002). Men who suffered from one or more of the above diseases were exposed to prostate cancer 6.5 times who did not suffer. Sutcliffe, S. et al. (2006) conducted a study about gonorrhea, syphilis, clinical prostatitis, and the risk of prostate cancer. They were asked participants to report their history of gonorrhea, syphilis, and clinical prostatitis by mailed questionnaire. Of the 36,033 participants in this analysis, 2,263 were diagnosed with prostate cancer. No association was observed between gonorrhea [adjusted relative risk "RR" was 1.04; 95% confidence interval "CI",(0.79, 1.36)] or syphilis [RR was 1.06; 95% CI,(0.44, 2.59)] and prostate cancer. They were also observed association between clinical prostatitis and prostate cancer (RR, 1.08; 95% CI, 0.96-1.20), although a significant positive association was observed among younger men (<59 years) screened for prostate cancer (RR, 1.49; 95% CI, 1.08-2.06; (p-(interaction) = 0.006). Miah, S. et al. (2014)

collected data from an online search and contemporary data presented at international urological congresses. They found a relationship between benign prostatic hyperplasia (BPH) and prostate cancer.

In this study, the effect of alcohol consumption on prostate cancer incidence was observed. Men who consumed alcohol were exposed to the disease 8.1 times than those who did not, Table (4.6.1). Brunner, Clair et al. (2017) found little evidence that variants in alcohol metabolizing genes were associated with prostate cancer diagnosis. The obesity or overweight was significantly related to the disease with χ^2- values 5.97 and 5.33, table (4.5.1) and (4.6.1) respectively. Men with obesity were more susceptible to the disease 3.6 times than normal men, with p- value less than 0.05 in both tables. Regarding with this point, a previous study (Snowdon, David A. et al. 1984) found that the obesity was a risk factor for cancer of prostate. Between 1960 and 1980 mortality data were collected from 6,763white men through a questionnaire on cohort members. Overweight men had a significantly higher risk of fatal prostate cancer than men near their desirable weight. The predicted relative risk of fatal prostate cancer was 2.5 for overweight men.

Intake of fruits and vegetables was important factor to decrease the risk of the disease incidence. This study showed that men who did not intake fruits and vegetables were more susceptible to the disease 3.8 times more than who did. The p-value was 0.000 and 0.0293 from Table (4.5.1) and table (4.6.1) respectively. Kirsh, Victoria A. (2007) evaluated the association between prostate cancer risk and intake of fruits and vegetables in 1338 patients with prostate cancer among 29361 men and Cox proportional hazards models were used. They demonstrated that intake of fruits and vegetables decreased the incidence of prostate cancer.

This study found association between animal fat intake and occurrence of prostate cancer. Men who did intake animal fat were exposed to the disease 4.1 times more than who did not. A previous study (Le Marchand et al. 1994) clarified that the role of animal fat on the incidence of prostate cancer and indicated that it may act by shortening the latency period of the disease.

Chapter 5

Conclusions & Recommendations

Chapter five

(5.1) Conclusions:

From the results of the study the researcher believes that it had achieved its objective, because it proved to us that:

- The most important predictive risk factors for prostate cancer incidence which determined by the logistic regression procedure were age, PSA and state=3; States of (The former Central Region in Sudan), Khartoum state as reference group. By comparing p-value with ($\alpha = 0.05$), there is no strong evidence to reject the null hypothesis. So, alcohol, state=2(the states of northern and eastern Sudan) and state=1(Darfur and kurdufan states) were insignificant. The percentage correctly classified was 98.4%., so the resulting model was appropriate and accurate.
- Chi-square test demonstrated that all variables were significantly associated with the outcome variable (diagnostic), except the smoking variable was insignificant.
- Mantel-Haenszel test showed that the following risk factors:(PSA, Age, marital status, diseases, alcohol, weight, intake of green vegetables , intake animal fat) were significantly associated with prostate cancer incidence (the outcome variable). Other variables that were not related to the outcome variable were: occupation, family history, cholesterol, blood pressure, intake of prostate medication and smoking with (p-value> 0.05), the variable of state had been excluded; it consists of 4 categories.
- The variables agreed by the Mantel-Haeszel test and the Chi-square test that affect on the incidence of prostate cancer shown above were (PSA, age, diseases, alcohol, weight, intake of green vegetables, animal fat and marital status). The smoking was insignificantly related to the disease as shown by both tests with p-value 0.937 and 0.587, Table1 and Table2 respectively. Chi-square test showed the other significant variables that related to the disease, whereas the Mantel-

Haenszel test showed that they were insignificant variables. These variables were (intake prostate medication, occupation, cholesterol, blood pressure and family history).

- The most important risk factors that agreed by all three procedures: logistic regression model, chi-square test and Mantel-Haenszel test were age and PSA.
- The chi-square test is the best in terms of determining the risk factors for the disease because it contains the highest χ^2 values for the variables.

(5.2) Recommendations:

Based on the research findings and discussion of the results, the following points are to be recommended:

1- There should be an optimum use of the Multiple Logistic Regression in designing statistical classification models or in group separation, especially when there is a mixture of variables between the continuous and discrete variables, or when the variables or do not follow a normal distribution.
2- Maximum use of the Mantel-Haenszel procedure in the biostatistics field because it is one of the most important statistical procedures that dealing with the stratified tables.
3- Maintaining a balanced diet, by reducing the animal fat and increase the intake of green vegetables and fruits.
4- Avoiding the obesity and maintaining an ideal weight to reduce the risk of prostate cancer.
5- Reducing the intake of alcohol because it contributes to the appearance of prostatitis which leads to cancer.
6- Avoiding all diseases that contribute to the incidence of prostate cancer, these diseases are: “syphilis, gonorrhea, chronic prostatitis and prostate enlargement”.
7- Raising awareness of the need to examine PSA periodically, especially when the age equal to or above 50 years, because age is the strongest risk factor for the appearance of prostate cancer.

8- More studies should be carried out in the former central region because there is a high prevalence of prostate cancer in this area.
9- Further research and studies need to be conducted about prostate cancer to investigate the most dangerous risk factors that increase the prevalence of the disease, because it affects a large group of people in the society.

References:

1- Adejumo, A. O. and Adetunji, A. A. (2013). Cochran–Mantel–Haenszel test for repeated tests of independence: An application in examining students' performance [online]. Journal of Education and Practice. **4**(23): 10-17. Available from <http://iiste.org/Journals/index.php/JEP/article/view/8366>

2- Adeloye, Davies and David, Rotimi Adedeji et al. (2016). An estimate of the incidence of prostate cancer in africa: A systematic review and meta-Analysis. PLoS One Journal. **11**(4): 1-18. doi:10.1371/journal.pone.0153496.

3- American Cancer Society 2016[a]. *Cancer facts and figures*. Accessed 13.Jul.2017. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2016/cancer-facts-and-figures-2016.pdf>

4- American Cancer Society 2016[b]. *Key Statistics for Prostate Cancer*. Accesseds date 10.July.2017. <https://www.cancer.org/cancer/prostate-cancer/about.html>

5- American Cancer Society 2016[c]. *Cancer facts figures for African Americans.* Accessed 20.June.2017. <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-facts-and-figures-for-african-americans/cancer-facts-and-figures-for-african-americans-2016-2018.pdf >

6- American Cancer Society 2016[d]. *What Causes Prostate Cancer: Acquired gene mutations.* Accessed 30.January.2017. <https://www.cancer.org/cancer/prostate-cancer/causes-risks-prevention/what-causes.html>.

7- American Cancer Society 2016[e]. *Prostate Cancer Risk Factors*. Accessed 30.January.2017. <https://www.cancer.org/cancer/prostate-cancer/causes-risks-prevention/risk-factors.html> .

8- American Cancer Society 2016[f]. *Prostate Cancer: What are the signs and symptoms of prostate cancer?* Accessed 30.January.2017.

<http://www.medicinenet.com/prostate_cancer/page4.htm#what_are_the_signs_and_symptoms_of_prostate_cancer >.

9- American Cancer Society 2016[g]. *Early Detection, Diagnosis, and Staging*. Accessed 10.March.2017. <https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging.html >.

10- American Cancer Society 2016[h]. *Prostate Cancer Screening Test Results Aren't Normal*. Accessed 11.March.2017. <https://www.cancer.org/cancer/prostate-cancer/early-detection/if-test-results-not-normal.html>.

11- American Cancer Society 2016[i]. *Prostate Cancer Stages: The AJCC TNM staging system*. Accessed 10.April.2017. <https://www.cancer.org/cancer/prostate-cancer/detection-diagnosis-staging.html>.

12- American Cancer Society 2016[j]. *Treating Prostate Cancer*. Accessed 12.April.2017. <https://www.cancer.org/cancer/prostate-cancer/treating.html>.

13- Barra, Mathias et al. (2014). Statistical testing of association between menstruation and migraine. Journal of Head and Face Pain. **55**(2):229-40. doi: 10.1111/head.12457.

14- Bolboacă, Sorana D. et al. (2011). Pearson-Fisher chi-square statistic revisited. The Information Journal. ***2***(3): 528-545. doi: 10.3390/info2030528.

15- Brunner, Clair et al. (2017). Alcohol consumption and prostate cancer incidence and progression: A Mendelian randomization study, Int J Cancer. **140**(1): 175–85. doi: 10.1002/ijc.30436.

16- Carter, H.B., Piantadosi, S., Isaacs J.T. (1990). Clinical evidence for and implications of the multistep development of prostate cancer, J Urol. **143**(4), 742-6.

17- Christensen, Ronald. (1997). Log-linear models and logistic regression. Second edition. Springer-verlag New York. Inc.

18- Cramer, Scott D. 2007. Deadly diseases and epidemics: prostate cancer. Chelsea House. New York. doi: 10.1016/S0002-9440(10)64691-3.

19- Ernesto, E.P., D., Giovanni, Jaileen, R. and Stephanie, L.M. (2016). An analytical study of prostate-specific antigen dynamics. Computational and Mathematical Methods In Medicine. **2016** (2016): 6 pages. Article id 3929163. doi:10.1155/2016/3929163.

20- Fidler, Vaclav and Nagelkerke, Nico. (2013). The Mantel-Haenszel procedure revisited models and generalizations. PLOS ONE Journal. **8**(3): 1-4. doi:10.1371/journal.pone.0058327.

21- Frank, Todd Michael et al. (2011). The chi-square test often used and more often misinterpreted. American Journal of Evaluation. **33**(3): 448-458. doi: 10.1177/1098214011426594.

22- Gacci, M. and Russo, G. I. et al. (2017). Meta-analysis of metabolic syndrome and prostate cancer. Prostate Cancer and Prostatic Diseases. **20**(2): 146-155. doi:10.1038/pcan.2017.1.

23- Gökce, M.I. et al. (2017). Is active surveillance a suitable option for african american men with prostate cancer? A systemic literature review. Journal of Prostate Cancer and Prostatic Diseases. **20**(2): 127–136. doi:10.1038/pcan.2016.56.

24- Goodness of Fit in Logistic Regression (2013).Accessed 12.February.2017. <http://www.medicine.mcgill.ca/epidemiology/joseph/courses/epib-621/logfit.pdf>

25- Heylighen, F. (1997). *Occam's Razor*. Accessed 10.August.2017. <http://pespmc1.vub.ac.be/OCCAMRAZ.html >.

26- Hosmer, David W., Lemeshow, Stanley (2002). Applied Statistics Regression. Second edition. Wiley, New York.

27- Jaret, Peter 2010. *Prostate Biopsy and the Gleason Score*. Accessed 26.March.2017. <http://www.webmd.com/prostate-cancer/features/prostate-biopsy-and-the-gleason-score#2>

28- Kanbayashi, Yuko et al. (2013). Predictive factors for agitation severity of hyperactive delirium in terminally ill cancer patients in a general hospital using ordered logistic regression analysis. Journal of Palliative Medicine. **16**(9): 1020-1025. doi: 10.1089/jpm.2013.0100

29- Kirsh, Victoria A., Peters, Ulrike, Mayne, Susan, T., Subar, Amy F., Chatterjee, Nilanjan, Johnson, Christine C. (2007). Prospective Study of Fruit and Vegetable Intake and Risk of Prostate Cancer. J Natl Cancer Inst. **99**(15): 1200-1209. doi:10.1093/jnci/djm065.

30- Le Marchand, Kolonel, Wilkens, Myers, Hirohata. (1994) Animal fat consumption and prostate cancer: a prospective study in Hawaii. Epidemiology **5**(3): 276-82.

31- Li, Shou-Hua , Simon, Richard, M. , Gart, John J. (1979). Small sample properties of the Mantel-Haenszel test, Biometrika. **66**(1): 181-183. doi: 10.1093/biomet 66.1.181

32- Ludbrook, John. (3013). Analyzing 2x2 contingency tables: 'which test is best?'. Journal of the Clinical and Experimental Pharmacology and Physiology. **40**(3): 177–180. doi: 10.1111/1440-1681.12052.

33- Mehta, Cyrus R., Patel, Nitin R. (1996).SPSS exact tests 7.0 for windows. SPSS Inc. United States of America.

34- Menard, Scott. (2002). Applied logistic regression analysis, second edition. SAGE publication. London, New Delhi.

35- Miah, S. and Catto, J. (2014). BPH and prostate cancer risk. Indian J Urol. **30**(2): 214–218. doi: 10.4103/0970-1591.126909.

36- Moosazadeh, M. et al. (2015). Predictive factors of death in patients with tuberculosis: a nested case–control study. Eastern Mediterranean Health Journals. **21**(4): 287- 292. doi: 10.1089/jpm.2013.0100.

37- Moreira, D.M. et al. (2017). The combination of histological prostate atrophy and inflammation is associated with lower risk of prostate cancer in biopsy specimens. Journal of Prostate Cancer and Prostatic Diseases, **20**(2): 1-5. doi: 10.1038/pcan.2017.30.

38- Mukhtar, Rayyan I., Hamdi, Ahmed M., Mohmmed, Omer A. (2017). Assessment of the potential risk factors of prostate cancer: A comparative study. Iosr-Jm. **13**(3): 41-45. doi: 10.9790/5728-1303024145.

39- Mukhtar, Rayyan I., Hamdi, Ahmed M., Mohmmed, Omer A. (2017). Predictive risk factors of prostate cancer incidence in Khartoum State, Sudan. Iosr-Jm. **13**(2): 42-46. doi: 10.9790/5728-1302044246.

40- NCSS, LLC. (2016). *Mantel-Haenszel Test*. Accessed 30.April.2016 <https://ncss-wpengine.netdna-ssl.com/wp content/themes/ncss/pdf/Procedures/NCSS/MantelHaenszel_Test.p df>

41- Nicholson, L. and Hotchin, H. (2015). The relationship between area deprivation and contact with community intellectual disability psychiatry. Journal of Intellectual Disability Research. **59**(5): 487–492. doi: 10.1111/jir.12149.

42- O'Connell, Ann A. (2006). Logistic regression models for ordinal response variables. First edition. SAGE publications, United States of America.
43- Onchiri, Sureiman. (2013). Conceptual model on application of chi-square test in education and social sciences. *Educational Research and Reviews*. ***8****(15): 1231-1241. doi: 10.5897/ERR11/0305.*
44- Pamela I. Ellsworth 2016. *Causes of prostate cancer*. Accessed 1.July.2016. <http://www.medicinenet.com/prostate_cancer/page2.htm#what_causes_prostate_cancer>
45- Peng, Chao-Ying Joanne et al. (2002). An introduction to logistic regression analysis and reporting. Journal of Educational Research. **96**(1): 3-14.
46- Prostate Cancer Foundation of Australia 2016. *What you need to know about prostate cancer*. Accessed 10.July.2017. <http://www.prostate.org.au/awareness/general-information/what-you-need-to-know-about-prostate-cancer/>
47- Rao, C.R. (2008). Epidemiology and medical statistics. First edition. Elsevier B.V. North-Holand.
48- Report of the Sudanese Federal Ministry of Health. 2016. Prostate cancer. Khartoum state.
49- Report of the Sudanese Federal Ministry of Health. 2016. Prostate cancer. Khartoum state.

50- Rosenbaum, Paul R. and Dylan, S. (2016). An Adaptive Mantel–Haenszel Test for Sensitivity Analysis in Observational Studies. Journal of the International Biometric Society. **73**(2): 422–430. doi: 10.1111/biom.12591.
51- Snowdon, David A., Philips, Ronald L., Choi, Warren, (1984). Diet, obesity, and risk of fatal prostate cancer, Am J Epidemiol. **120**(2): 244-250. doi: 10.1093/aje.a113886.
52- Sutcliffe, S., Giovannucci, E., De Marzo, A.M., Leitzmann, M.F., Willett, W.C., Platz, E.A. (2006). Gonorrhea, syphilis, clinical prostatitis, and the risk of prostate cancer. Cancer Epidemiology, Biomarkers & Prevention. **15**(11): 2160-6. doi: 10.1158/1055-9965.
53- Swanson, Kristin R., True, Lawrence D., Lin, Daniel W., Buhler, Kent R., Vessella, Robert, and Murray, James D. (2001). A quantitative model for the dynamics of serum prostate-specific

antigen as a marker for cancerous growth, Am J Pathol. **158**(6): 2195–2199.

54- Tripepi, Giovanni et al. (2010). Stratification for confounding – part1: the mantel-haenszel formula. Nephron Clinical Practice Journal. **56**(1): 129–140. doi: 10.1159/000319590.

55- Tyson, M.D.1., Andrews, P.E., Etzioni, D.A., Ferrigni, R.G., Humphreys, M.R., Swanson, S.K., Castle, E.K. (2013). Marital status and prostate cancer outcomes. Can J Urol. **20**(2): 6702-6.

56- Vidal, A.C. and Howard, L.E. et al. (2017). Obesity and prostate cancer-specific mortality after radical prostatectomy. Journal of Prostate Cancer and Prostatic Diseases. **20**(1): 72-78. doi:10.1038/pcan.2016.47.

57- Wayne, W. Daniel. (2009). Biostatistics, A Foundation for Analysis in the Health Sciences. Ninth edition. John Wiley & Sons, Inc, United States of America.

58- Weiner, A.B. et al. (2017). Contemporary management of men with high-risk localized prostate cancer in the United States, Journal of Prostate Cancer and Prostatic Diseases. **20**(2): 1-6. doi: 10.1038/pcan.2017.5.

59- Weisberg, Sandford . (2005). Applied linear regression. Third edition. USA. Published online.
< http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf >.

60- Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. Journal of the Royal Statistical Society **1**(2): 217–235.

61- Zhang, B. et al. (2011). Assessment of risk factors for early seizures following surgery for meningiomas using logistic regression analysis. Journal of International Medical Research. **39**(5): 1728 – 1735. doi:10.1177/147323001103900515

Appendixes:

This questionnaire aims to identifying the risk factors for prostate cancer incidence. Please complete the form carefully. The information we obtain is confidential and is for scientific research only:

Tick (√) in the box with the appropriate answer

1- Age :

☐ Less than 50 ☐ greater than or equal 50

2- Occupation:

☐ Jobs depend on prolonged seating

☐ Jobs depend on physical activities

3- State:

☐ Darfur and kurdufan states

☐ States of (previously central region)

☐ The states of Northren and Eastern Sudan

☐ khartoum state

4- Marital status:

☐ Married ☐ Unmarried

5- Family history :

☐ Yes ☐ No

6- Red meat and animal fat:

☐ Yes ☐ No

7- Vegetables and fruits:

☐ Yes ☐ No

8- Overweight:

☐ Yes ☐ No

9- High cholesterol:

☐ Yes ☐ No

10- High blood pressure :

☐ Yes ☐ No

11- Intake of Prostate medication:

☐ Yes ☐ No

12- Alcohol:

☐ Yes ☐ No

13- Smoking:

☐ Yes ☐ No

14- Developing one or more of these diseases: 'Syphilis, gonorrhea, chronic prosatitis, and prostate enlargement':

☐ Yes ☐ No

15- Prostate Specific Antigen (PSA):

☐ Normal ☐ Abnormal