MPH Dissertation

Epidemiology of oesophageal cancer at Groote Schuur Hospital.
A case control study

by
Dorina Saleh
SLHDOR001

Abstract :

Cancer of the oesophagus is one of the most common cancers in South Africa. It occurs more predominantly among men but is not completely absent among the women. For this study we collected demographic information on women with oesophageal cancer being treated at Groote Schuur Hospital in Cape Town, South Africa. The data formed part of a general cancer database in which incident cases of cancer confirmed by histopathology were collected. The majority of the patients originated from Eastern Cape province of South Africa and lived in rural areas. No data on socio economic status was collected but data on fuels used for cooking, race and smoking status were collected. It emerged that women who used wood or paraffin for cooking had a higher risk for developing this cancer than those who use electricity. The OR were 2.8(95% CI 1.32-5.92) for wood and 2.9(95%CI 1.19-7.051) for paraffin.

Smoking was also found to contribute significantly to the risk of developing oesophageal cancer with OR 3.24 (95%CI 1.49-7.04). It also appeared that paraffin and wood contributed more to the risk in rural black women and smoking in rural coloured women. This is the first report of an association between fuels types and oesophageal cancer in women from the Eastern Cape of South Africa. Although smoking has previously shown to be associated with lung cancer in women this will be the first report of an association between smoking and oesophageal cancer in coloured women in the Western Cape.

Index

1. Introduction:

Cancer of the oesophagus is the 8th most common cancer in the world, but its importance varies greatly from country to country. There are several hot spots for oesophageal cancer in the world; these include areas in the northern Iran, the northern (Linxian district) China and South Africa where annual incidence can exceed 100 per 100000 and over 20% of deaths are associated with oesophageal cancer. North America and Europe have much lower incidences. These range between 5-10 cases per 100000 annually[11].

Oesophageal cancer is a disease of middle to late adulthood, rarely occurring in persons under 25 years of age and occurring more frequently in men than women [7]. There is also a strong association with racial differences with oesophageal cancer occurring more frequently among black peoples than whites [3]. The excess of oesophageal cancers in blacks is greater among individuals younger than 55 compared to their white counterparts [7].

We distinguish mainly two histological subtypes of oesophageal cancer: squamous cell carcinoma and adenocarcinoma [1]. More than 90% of oesophageal cancers can be categorised as either of these two subtypes [2].

Among black people and in most developing countries, squamous cell carcinoma is the most prevalent type and adenocarcinomas occur predominantly among whites [1, 2]. The main identified risk factors for squamous cell carcinomas are agents that are likely to cause irritation and inflammation to the oesophageal mucosa such as cigarette smoke and alcohol[1, 2].

In the US and a western country such as France, where adenocarcinomas is the more common type, the main risk factors are cigarette smoking and alcoholic beverage consumption.

The risk increases with increasing amount of alcohol consumed [7, 14]. Numerous case-control studies have shown that both drinking and smoking increase the risk of oesophageal cancer in a synergistic manner [7]. In India, risk of oesophageal cancer has been reported as higher among chewers of quids containing tobacco and betel [16]. Drinking of specific alcoholic beverages has been implicated in several clusters of increased oesophageal cancer mortality around the world; apple brandies in France, maize beers for South Africans and moonshine whiskey in South Carolina, USA [7].

Risk of squamous cell oesophageal cancer has been shown to be higher in populations where limited diets prevail, particularly diets deficient in fresh fruits and vegetables [17]. A case control study conducted in the Linxian district of China found a slight increased risk with higher intake of fresh vegetables and legumes [18]. This is contrary to most findings that prescribe supplementation of these elements in the diet to reduce the risk of oesophageal cancer. Carotenoids and vitamin C and E are thought to be involved as inhibitors of oesophageal cancer [7, 18]. Other dietary elements that have been implicated in the risk of oesophageal cancer in Sweden are heterocyclic amines [9].

Heterocyclic amines are carcinogenic substances that are formed through pyrolysis of amino acids and creatine or creatinine when meats are cooked at high temperatures, particularly by pan-frying. The contents of heterocyclic amines increase with increasing cooking temperature [9]. There seems to be reason to suspect heterocyclic amines in the aetiology of squamous cell oesophageal cancer but not adenocarcinomas [9].

On the whole, the risk of squamous cell oesophageal carcinoma has been strongly associated with socio-economic status with its predominant occurrence in developing countries and among poor people.

However, the presence of hot spot within developing countries such as South Africa and the disparities between the incidence of this cancer between poor whites and poor black Americans suggests that there may be more than just socio economic status to this association. A recent case control study in Scotland found no clear association between deprivation and risk of oesophageal adenocarcinoma and cancer of the gastric cardia [15], while an American survey of mortality data found evidence that suggests that poor white Americans have lower incidences of adenocarcinoma than their wealthier counterparts and an overall lower incidence of squamous cell carcinoma relative to poor African Americans [8]. Very limited information is available regarding risk factors for the development of oesophageal cancer with respect to racial disparities.

This may indicate some form of genetic predisposition in some population groups to develop certain subtypes of oesophageal cancer. However, this seems to be disputed by the results of case control study in Sweden showing no association between family history of oesophageal cancer and risk of the cancer, regardless of histological type [10]. This is in contrast with gastric cardia cancer patients, who reported a family history of gastric cancer more often than the control subjects.

It has been hypothesised that the origins of oesophageal and gastric cardia adenocarcinomas are similar to one another and distinct from squamous cell carcinomas. Indeed Oesophageal adenocarcinomas appear to have a different aetiology but information on risk factors for adenocarcinomas of the oesophagus is scanty.

Like squamous cell carcinoma, adenocarcinomas of the oesophagus are also associated with increased consumption of alcoholic beverages and smoking, although the association is somewhat lower than for squamous cell carcinomas [14].

The strongest association have been found with recurring symptoms of reflux or factors that have been shown to be associated to recurring symptoms of reflux such as drugs that relax the gastro-oesophageal sphincter, obesity and Barrett's oesophagus [2].

Barret's oesophagus is the condition in which metaplastic columnar epithelium replaces the squamous epithelium in the distal oesophagus. It seems to be a result of chronic gastro-oesophageal reflux disease (GERD) [19]. Barrett's oesophagus is the single most important risk factor for oesophageal adenocarcinoma. Like adenocarcinoma of the oesophagus, Barrett's oesophagus occurs more frequently in white people than in blacks [1, 7].

In South Africa, oesophageal cancer has a particularly high incidence, especially among the black population [3]. It is currently the commonest cancer in black males [5] and thus is an important public health concern. It is most prominent in people originating from South Transkei, Eastern Cape [20]. It has been found that district hospitals in Transkei report more oesophageal cancer cases relative to other areas of South Africa[5]. In a survey on time trends of incidence of cancers in South African gold miners, of the cancer patients who originated from the Transkei region, over half of them had oesophageal cancer and predominantly squamous cell carcinomas [12]. In the Transkei region, males are more frequently affected than females. This however is not apparent in Ciskei [7].

In addition to smoking and alcohol intake, in South Africa the excess risk is suspected to come from maize derived beers and tobacco, especially pipe tobacco. Present evidence suggests that nutritional deficiencies may also play a part [6].

Low levels of fresh fruits and vegetables in the diet and the consumption of *Fusarium*-contaminated foods have been shown to correlate with higher incidence of squamous cell-type oesophageal cancer[1]. *Fusarium verticillicoides* is a fungus that attacks maize; these synthesize fumonisins, which have been shown to have tumour-promoting activity [1].

Although there is much information about risk factors of oesophageal cancer in the Transkei region, there is dearth of epidemiological data for other parts of South Africa. It is possible that these differ for other cohorts. Thus it is important to look at other areas of South Africa and determine if the risk factors differ across the country and maybe begin to give an explanation of why people in Transkei have such an elevated risk compared to other parts of South Africa.

This study was done to probe for risk factors in oesophageal cancer patients that receive treatment at Groote Schuur Hospital (GSH) in the Western Cape. We sought to bring new epidemiological data on oesophageal cancer and look at the rural to urban differential and the effects of the use of different fuels for cooking on the risk of oesophageal cancer the Western Cape province, particularly at the population served by GSH.

GSH is a tertiary care facility in the Cape Town metropolis and serves as a referral centre for a large part of the Western Cape. Thus the main "risk factors" looked at are whether the participants live in rural or urban areas, their gender, age, smoking status and the sources of energy that they used for cooking.

2. Objectives:

The first objective of this study is to determine whether the following factors are important predictors of occurrence of oesophageal cancer in black and coloured patients that are treated at GSH:

- Locality of the place of birth
- Locality of the current place of residence
- Age
- Gender
- Sources of energy used for cooking:
  - 20 years ago
  - Currently
- Smoking status

The second objective is to describe the demographic characteristics of black and coloured oesophageal cancer patients at GSH.

3. Methods

3.1 Study design

This is a hospital based case control study in which cases are oesophageal cancer patients (cases) admitted at GSH and controls are patients that were admitted for cancers that are thought to be unrelated to oesophageal cancer.

The study population is all black and coloured people over 18 years of age, admitted at GSH with a diagnosis of cancer. Only incident cases of primary invasive cancer diagnosed by histology, cytology or haematology in black and coloured patients were included in the study. Participants were recruited with cooperation of oncologists, surgeons and pathologists at GSH.

Demographics and risk factor data was collected by nurses trained to administer validated questionnaires. Interviewers sought out patients with new diagnoses by visiting each of the outpatient, surgical and radiotherapy wards. The patient's diagnosis, histological typing, stage and ICD code were recorded in a cancer database at GSH.

The cases were all diagnoses with ICD-10 classification C15.3 and 15.9 are malignant neoplasms of the oesophagus. Controls were all diagnoses with ICD-10 (1989) codes 16.9, 18.7, 19.5, 21.1 and 21.8 which relate to neoplasm of different parts of the digestive system, C34.2 neoplasm of the bronchus and lungs, C50.9 breast cancer, C51.9, 52.9, 53.9, 54.1, 54.9, 55.9 and 56.9 neoplasms of various areas of female reproductive organs and C80.9 unspecified neoplasms.

These codes were interpreted according to the 10th revision of International Classification of Diseases (ICD-10) 1989 guide [21]. The cases comprised of both adenocarcinoma and squamous cell carcinoma of the oesophagus.

In total, at the outset, we had 475 participants recorded; 241 cases and 234 controls but 23 were removed due to missing data.

3.2 Statistical analyses

Multivariate logistic regression analysis was used to model the relationships between risk of oesophageal cancer and the risk factors of interest and to estimate the relative risks of developing oesophageal cancer.

This was a case control study thus only OR can be calculated. Incidence rates and prevalence of oesophageal cancer were not determined as the source of cases and controls does not allow the determination of the size of the baseline population. A significance level of 5% was used.

Whereas the chi-squared and t tests indicate whether or not there is an association between the variables and outcome of interest, logistic regression affords us the ability to explore the nature of that association.

The source of energy for cooking was a categorical variable with more than two categories thus in order to effectively determine the OR for each fuel. I created dummy variables or sub variables within a risk factor. The creation of dummy variables allows us to compare the risk of oesophageal cancer relative to different categories, whether these are ordered or not, within each risk factor.

The reference category is used as the baseline or comparison group for all the other categories, so that the quoted OR are the approximated relative risk of disease for patients in a particular category relative to those in the reference category of each risk factor.

Initially logistic model relating individual risk factors to the occurrence of oesophageal cancer were fitted in order the find the risk factors with the significant associations to the outcome. I then fitted regression models that between the risk factors that showed significant relationship with occurrence of oesophageal cancer. This was done so as to determine if they were associated to each other and this way identify possible confounding factors. Finally the model was built by forward selection and possible confounding variables were entered first. The rest of the variables were considered each in turn to see which combination of risk factors gave the best prediction given the variables already in the model. The models were then compared on the basis of the log likelihood of the model using the log likelihood ratio test (Lrtest) for nested models and the Aikaikes information criterion (AIC) for non-nested models.

The final and best model was determined on the basis of the Aikaike's information criterion (AIC) value and log likelihood values.

Logistic models are fitted by maximum likelihood estimation to estimate ß coefficients that estimate the maximum likelihood of occurrence of sample results. Thus, the higher the log likelihood of a model, the better the model. The likelihood ratio test essentially compares the deviances of two models provided that one is nested in the other. The AIC value is used to compare two models that are not nested, the model with the lowest AIC is the better model. A summary of the model building process and model selection is found in the appendix. Table 6 depicts the final model from that process.

The calculated odds ratios (OR) give an approximation of relative risk of oesophageal cancer for each category of the risk factors

Outlying observations were identified on the basis of the values of their residuals. In this case both the standardised deviance (ds) and pearson (rs) residuals were generated. An entry with a residual value greater than 2 is considered as an outlier. The h, a measure of leverage was generated. A value significantly greater than 2p/n, essentially 0 is considered to be influential. The d2 value measures the change in the pearson goodness of fit statistic with or without a particular covariate pattern in the model. A relatively large value is indicative of high influence. Any individual with a d2 value over 10, I considered influential. Potentially influential observations were removed from the model and the final model was fitted again to see whether the odds ratio changed significantly as a result.

3.3 Ethical consideration

All information pertaining to the identity of the participants were not entered in the database, thus ensuring their confidentiality. The study was not in any way harmful to the participants and hopefully the knowledge gained from the outcomes of the study will help in exploring ways of preventing oesophageal cancer.

The study has received ethical approval from the University of Cape Town (UCT).

4. Results

4.1 Data exploration

The final sample had a total of 223 controls and 229 cases. The cases were individuals diagnosed with oesophageal cancer (ICD 15.3 and 15.9) and controls were individuals diagnosed with a variety of cancers (ICD 16.9 to 80.9 in table 1).

Table1: Cancer Histopathology and corresponding ICD-10 codes in original sample

ICD code	Freq.	Percent	Description
Cases			
15.3	1	0.22	Upper third of the oesophagus
15.9	228	50.44	Oesophagus, unspecified
Controls			
16.9	6	1.15	Stomach, unspecified
18.7	2	0.44	Sigmoid Colon
18.9	3	0.66	Colon, unspecified
19	5	1.11	Rectosigmoid junction
21.8	2	0.44	Overlapping lesions of the rectum, anal canal anus
34	2	0.44	Middle lobe, bronchus and lung
50.9	2	0.21	Breast, unspecified
51.9	7	1.55	Vulva, unspecified
52.9	3	0.66	Malignant neoplasm of the vagina
53.9	140	30.97	Cervix uteri, unspecified
54.1	10	2.21	Endometrium
54.9	25	5.54	Corpus uteri, unspecified
55.9	2	0.21	Uterus, unspecified
56.9	15	3.32	Ovary
80.9	2	0.44	Malignant neoplasm without specification of site

Table 2: Description of the histological classifications of the cases.

Histopathology	Frequency	%
Squamous cell Ca	210	91.7
Adeno-carcinoma	10	4.37
Basaloid Ca	1	0.44
Carcinoma in situ	2	0.87
Mucoepidermoid Ca	1	0.44
Oat cell Ca	1	0.44
Seromucinous Ca	1	0.44
Severe dysplasia Ca	1	0.44
Small cell Ca	2	0.87
Total	229	

Of the cases, 210 or 91.7% were various forms of squamous cell carcinomas, 10 (4.4%) were adeno-carcinomas and the rest were other classifications of oesophageal cancer (table 2).

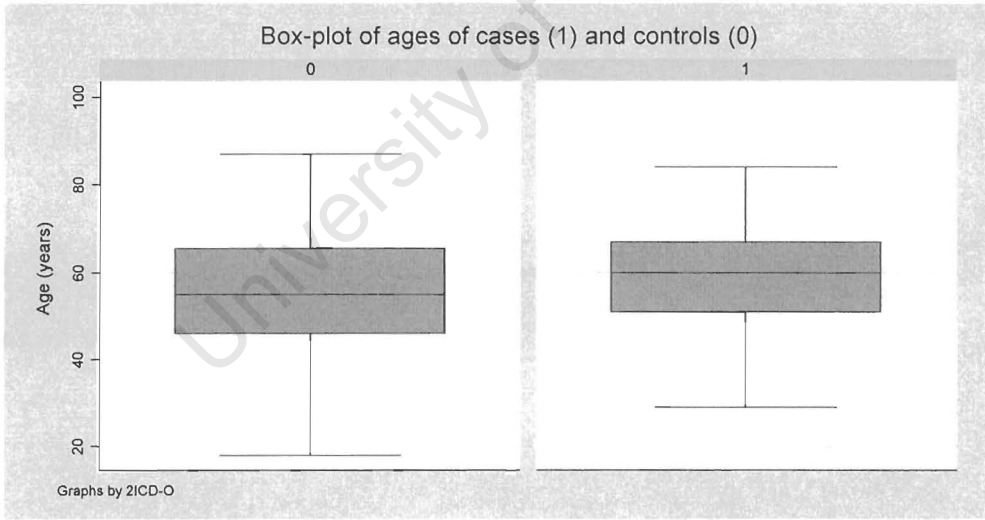
Table 3: Current province and province of birth of cases and controls

	Control subjects (n = 223)		Case subjects (n = 229)	
Province	No.	%	No.	%
Birth				
Eastern Cape	**102**	**45.74**	**139**	**60.70**
Western Cape	**99**	**44.39**	**75**	**32.75**
Free state	1	0.45	2	0.87
Gauteng	5	2.24	2	0.87
Kwazulu Natal	3	1.35	2	0.87
Northern Cape	1	0.45	1	0.44
Northern Prov	4	1.79	5	2.18
Other	8	3.59	3	1.31
Current				
Eastern Cape	**45**	**20.18**	**48**	**20.96**
Western Cape	**170**	**76.23**	**178**	**77.73**
Gauteng	1	0.45		
Kwazulu Natal	1	0.45		
Northern Province	3	1.35	2	0.87
Other	3	1.35	1	0.44

Patients were asked about their province of birth and the province that they currently call home. The data on birthplaces is shown in table 3. Most of the patients were born and currently live either in Eastern Cape or Western Cape. The proportions for patients who were born in Eastern and Western cape are 45.74% and 44.39% respectively for the controls and 60.70% and 32.75% for the cases. Many more cases originated from Eastern Cape than controls.

The proportions for patients who currently live in Eastern and Western cape are 20.18% and 76.23% respectively for the controls and 20.96% and 77.73% for the cases. These ratios are not significantly different between cases and controls. Many of the patients who indicated that they were born in Eastern Cape now reside in the Western Cape.

Graph 1: Box plot of ages of cases and controls

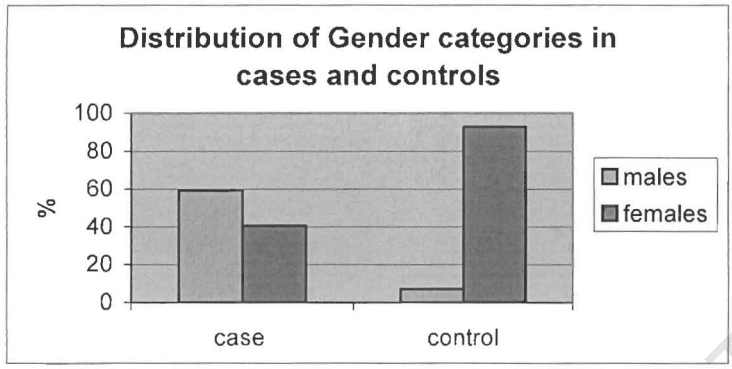


Shapiro-Wilk W test for normal data
Cases: P = 0.43
Controls: P = 0.30

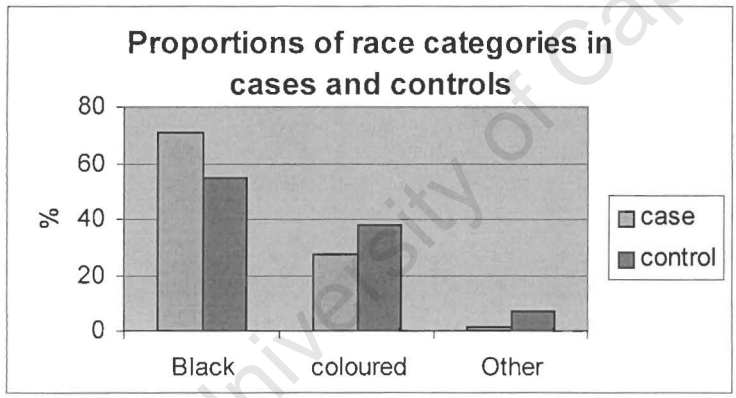
Graph 1 depicts the distribution of ages of cases and controls. Shapiro-Wilk test for normal data compares each distribution to a normal distribution. The importance of this test is that it helps to determine the correct descriptive measures to use for age.

For both cases and controls the test reveals a high probability that the age samples were drawn from normally distributed populations. Therefore the mean and standard deviation are appropriate descriptive measures.

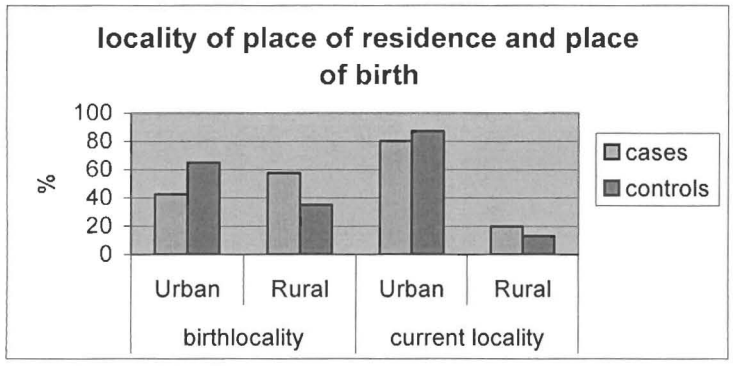
Graph 2



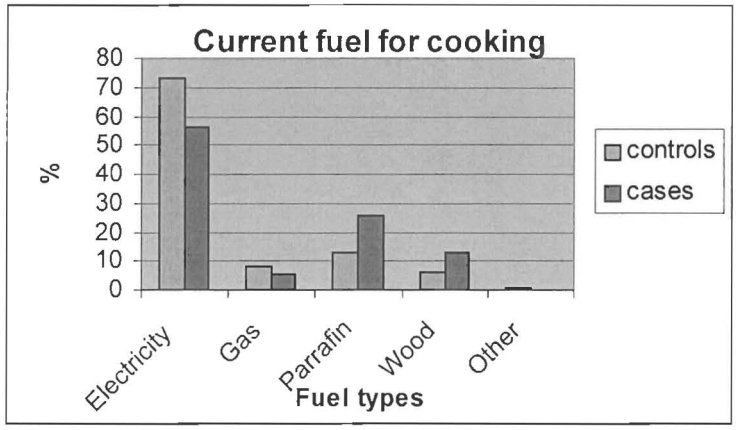
Graph 3



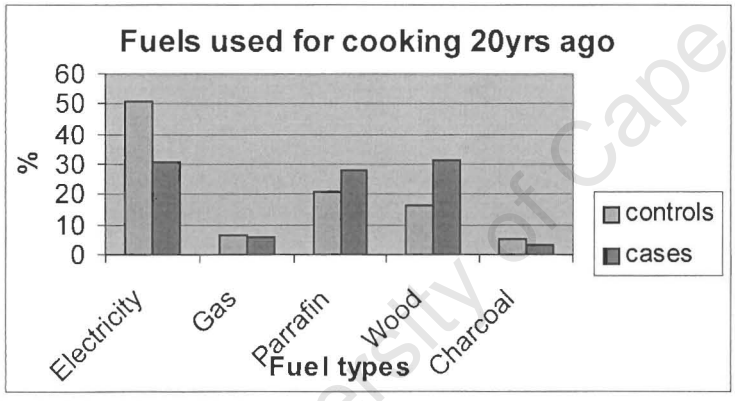
Graph 4



Graph5



Graph 6



Graph 7

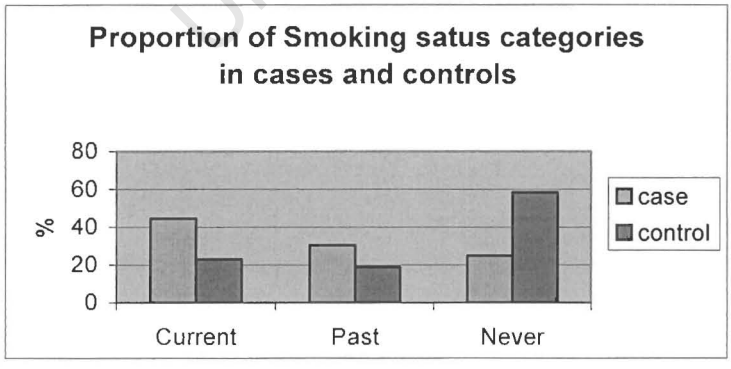


Table 4 contains summaries of the crude relationships of the individual risk factors and the occurrence of oesophageal cancer. Odds ratios (OR) were generated by fitting logistic regression models for each individual risk factor with occurrence of oesophageal cancer. The frequencies of cases and control for the variables of interest are found in table 4 and graphical representation of these are found in graphs 1 to 7. The explanatory variables being investigated are the ages of the patients, their race, gender, smoking status, locality of their places of birth and current places of residence, the sources of energy used for cooking currently and 20 years ago.

Age

The risk of oesophageal cancer is related to every one-year increase in age. The P value of 0.003 indicates significant OR. In this case the OR indicates the rate of change of the risk of oesophageal cancer with a one-year increase in age. Thus, when considering age on its own, the risk of oesophageal cancer increases by 2.2% with every yearly increase in age.

Gender

From the onset one can see that the male to female ratio is very different between the cases and controls (graph 2). There are 12.9 times more women than men in the control group relative to 0.68 in the cases. This indicates that the control sample is inadequate. From this sample it emerges that men are 18 times more at risk of developing the cancer than women.

Race

The race distribution between the cases and controls look relatively similar although there are about more 'coloured' people in the control than there are in the cases and more black people in the cases than the controls (graph 3).

People in the "Other" were not considered as missing data as people who are unwilling or unable to categorise themselves by race could potentially be different from those who do, and may have a different risk profile. People who categorised themselves as coloured have 44% reduced risk of oesophageal cancer than those who categorised themselves as black. People in the "Other" category have 81% reduced risk than the black people in the sample.

Locality of place of birth

People who are born in urban areas have a 60.5% reduced risk of developing the cancer than those born in rural areas and people who live in urban areas have a 39% reduced risk compared to rural dwellers.

Fuel for cooking

Patients were asked about the source of energy or fuel that they mostly used for cooking 20 years ago and currently. The predominant current sources of cooking energy are electricity, gas, paraffin and wood (graph 5). The majority of both the cases and controls used electricity. As far as energy sources go, people who use paraffin and wood as a fuel for cooking are at significantly greater risk of developing the cancer than those who use electricity.

This risk is increased by a factor of 2.66 and 2.82 for currently using paraffin and wood respectively and 2.25 and 3.14 respectively for using these fuels 20 years ago.

Table 4: Frequencies of cases and controls and crude OR for individual risk factors

Characteristics	Control subjects (n = 223) No.	%	Case subjects (n = 229) No.	%	OR	P value	95%CI
Age							
Mean age	55.74		59.24				
Std. Dev	13.71		11.08				
95% C I (mean)	(53.93 - 57.55)		(57.79 - 60.68)		1.02	0.003	(1.008 - 1.038)
Gender							
Female	207	92.83	93	40.61	(ref. Cat.)		
Male	16	7.17	136	59.39	18.9	0.000	(10.67 - 33.55)
Race							
(R1) Black	122	54.71	162	70.74	(ref. Cat.)		
(R2) Coloured	85	38.12	63	27.51	0.56	0.004	(0.37 - 0.83)
(R3) Other	16	7.17	4	1.75	0.19	0.003	(0.06 - 0.58)
Birth locality							
Rural	78	34.98	132	57.64	(ref. Cat.)		
Urban	145	65.02	97	42.36	0.395	0.000	(0.27 - 0.58)
Current locality							
Rural	29	13.00	45	19.65	(ref. Cat.)		
Urban	194	87.00	184	80.35	0.61	0.058	(0.37 - 1.02)
Energy now							
Electricity	163	73.09	129	56.33	(ref. Cat.)		
Gas	18	8.07	12	5.24	0.84	0.661	(0.39 - 1.81)
Paraffin	28	12.56	59	25.76	2.66	0.000	(1.61 - 4.41)
Wood	13	5.83	29	12.66	2.82	0.003	(1.41 - 5.64)
Other	1	0.45					
Energy 20 years ago							
Electricity	113	50.67	70	30.57	(ref. Cat.)		
Gas	14	6.28	13	5.68	1.50	0.328	(0.67 - 3.38)
Paraffin	46	20.63	64	27.95	2.25	0.001	(1.39 - 3.64)
Wood	37	16.59	72	31.44	3.14	0.000	(1.91 - 5.16)
Charcoal	12	5.38	8	3.49	1.29	0.608	(0.49 - 3.43)
Other	1	0.45	2	0.87			
Smoking status							
Current	51	22.87	102	44.54	4.56	0.000	(2.89 - 7.21)
In the past	42	18.83	70	30.57	3.8	0.000	(2.32 - 6.22)
Never	130	58.30	57	24.89	(ref. Cat.)		

Ref. Cat : Reference category

Smoking status

The majority of the controls have never smoked while the majority of the cases are current smokers (graph 7). Smoking emerges as a significant predictor of occurrence of oesophageal cancer. Those participants who currently smoke are 4.56 times more at risk than those who have never smoked and those who are no longer smokers but smoked in the past are 3.8 times more at risk than those who have never smoked.

From the exploratory results in table 4, it appears that there is significant relationship between the occurrence of oesophageal cancer and the age of patients, the race, the gender, location of the place of birth, the source of energy for cooking and smoking. There are however strong possible confounding variables. Logistic regression was used to explore these relationships because of the multiple categories within each risk factor.

4.2 Interrelationships between risk factors

In the case of the ages of patient, as age is a continuous variable I therefore fitted a linear regression model. Older people are at higher risk of developing oesophageal cancer. The assumption is that in addition to the fact that oesophageal cancer is a disease of old age, it is possible that an additional risk lies in the fact that those who are exposed to carcinogen emitting fuels and possibly unknown environmental risk factors are older. It emerged that individuals who smoked in the past are on average 4.92 years older than those who have never smoked and current users of wood for cooking are on average 5.6 years older than those who use electricity. In addition to these relationships, those who live in urban areas are on average 3.7 years younger than rural dwellers.

Table5: Important Inter-relationships between risk factors

	Smoke	Birth locality	Current locality	Energy now	Energy-20
Age coef. P value 95%CI	**In the past** 4.92 0.001 (2.01; 7.82)		-3.70 0.020 (-6.82; -0.58)	**wood** 5.16 0.012 (1.13; 9.19)	
Gender OR P value 95%CI	**Currently** 15.06 <0.05 (7.89; 28.75) **In the past** 14.38 <0.05 (7.32; 28.22)		2.46 0.004 (1.33; 4.57)	**wood** 0.18 0.001 (0.06; 0.52)	**wood** 0.55 0.032 (0.32; 0.95)
Race OR P value 95%CI	**Currently** 0.28 <0.05 (0.17; 0.44) **In the past** 0.61 0.059 (0.37; 1.02)	0.077 <0.05 (0.05; 0.13)	0.075 <0.05 (0.02; 0.21)	**Gas** 3.1 0.008 (1.3; 7.2) **Paraffin** 48.08 <0.05 (11.6; 199) **Wood** 22.6 <0.05 (5.37; 95.4)	**Gas** 3.8 0.002 (1.62; 8.97) **Paraffin** 11.19 <0.05 (6.1; 20.6) **Wood** 21.16 <0.05 (10; 44.7)
Smoke OR P value 95%CI		2.00 0.002 (1.29; 3.12)	2.45 0.008 (1.27; 4.72)		

OR: Odds Ratio
Coef.: linear regression β coefficient
P value
95%CI: 95% Confidence interval for the Odd Ratio or Coefficient

Women mainly do the cooking. It is therefore possible that fuel used for cooking is a stronger risk factor for women than men. This assumption is shown to be true. The results indicate that current users of wood for cooking are 82% more likely to be women and those who used it 20 years ago are 45% more likely to be women (table 5). The results in table 4 indicate that men are 18.9 times more at risk of oesophageal cancer than women and that users of wood are at increased risk of developing the cancer.

This means that exposure to wood is a stronger risk factor for women than it is for men. Also, men are 2.46 times more likely to reside in urban areas than women, which would explain why they are less likely to use wood for cooking. Moreover, men are 15 times more likely to be current smokers and 14 times more likely to be have smoked in the past than women (table 5). These indicate that smoking status may be a stronger risk factor for men than it is for women although this OR may be inflated because of the overwhelming number of women in the controls.

In the case of race it emerges that current smokers are 72% less likely to be black than coloured or of unspecified race category. This is only 39% for ex-smokers. Black people are also more likely to use gas, paraffin and wood. This by 3.1, 48.8 and 22.6 for current use of gas, paraffin and wood respectively and 3.8, 11.19 and 21.16 for using these fuels in the past. Moreover, black people are more likely to have been born (92.3%) and live (92.5%) in rural areas (table 5).

Smoking increases the risk of oesophageal cancer but these results indicate that it is probably a stronger risk factor for coloured people than it is for blacks and that whatever environmental risk factors are at play in rural areas, they probably impact more on black people that they do in coloured people.

In the case of smoking, current smokers and those who smoked in the past were most likely born in rural areas and are more likely to reside in rural areas, this by factor of 2 and 2.45 respectively (table 5). Since most smokers in this cohort are coloured, these results indicate that coloured people who smoked or smoke and at higher risk of oesophageal cancer, mainly originate and dwell in rural areas. Thus even in coloured populations the risk of oesophageal is probably related to behaviour trends in that are associated with low socio-economic background in the coloured population.

4.3 Multivariate logistic regression

The OR quoted up until now relate individual risk factors to the occurrence of oesophageal cancer but the results in table 5 indicate that there are interrelations between these factors that can influence the interpretation of the OR. The way to deal with this is to fit all these risk factor in a model that takes into account these relationships and adjusts for them thus improving our ability to predict the risk of oesophageal cancer.

Table 6: Significant risk factors in the final model

Variables	Reference category:	OR	95% CI		P value
Gender	Women	25.29	12.42	51.5	<0.05
birthlocal	Rural	0.38	0.21	0.69	0.002
age		1.028	1.01	1.05	0.010
Paraffin (C)		2.48	1.23	5.03	0.011
Wood (C)		3.19	1.35	7.54	0.008
Smoking (C)	Never smoked	3.48	1.76	6.86	<0.05
Coloured	Black	0.47	0.24	0.97	0.042
Unspecified race		0.16	0.03	0.80	0.025

The final model describes the following significant relationships:

Gender emerges as a very strong predictor for oesophageal cancer with an OR of 28.29 indicating that males are 28.29 times more at risk of developing the cancer than females. This is probably spurious as the controls were not adequate. The location of place of birth is also an important predictor of risk. People who were born in rural areas are 62.2% more at risk of developing oesophageal cancer than those who were born in urban areas.

The risk of oesophageal cancer increases by 2% with every year increase in age.

People who currently cook with paraffin have 2.48-fold increased risk of disease than those who use electricity.

Those who use wood have a 3.19 fold increased risk of disease than those who use electricity. People who currently smoke have a 3.48 fold increased risk of disease than those who never smoked in the past.

People who categorise themselves as coloured have a 53% reduced risk of developing oesophageal cancer than black people in the cancer ward at GSH. People who would not categorise themselves as either black or coloured (R3) have an even lower risk, 84% reduced risk, than their black counterparts.

These relationship appear unchanged when we consider only women in the sample (table 7).

Table 7 : Final logistic regression model applied to just women in the sample

Variables	Reference category:	OR	95% CI		P value
Birthlocal	Rural	0.34	0.17	0.66	0.002
Age		1.035	1.01	1.06	0.005
Paraffin (C)		2.80	1.32	5.92	0.07
Wood (C)		2.9	1.19	7.051	0.019
Smoking (C)	Never smoked	3.24	1.49	7.04	<0.05
Coloured	Black	0.46	0.20	1.06	0.069
Unspecified race		0.24	0.03	2.09	0.197

The same observations that were potentially influential in the model for the whole sample were also shown to be potentially influential in the model with just women. A sensitivity analysis was performed by removing potentially influential observations for the sample and refitting the final model for the just women. The OR did not change significantly.

5. Discussion and conclusion

Groote Schuur Hospital is a tertiary hospital in the heart of Cape Town and as such receives patients mainly from the western Cape. From this study it emerges that many of the patients who use the public service of GSH either came directly from eastern Cape or originated from Eastern Cape. Significantly more cases originated from Eastern Cape than controls. Moreover, significantly more cases were born in rural areas than controls.

This may be significant in that the Eastern Cape is one of the least developed and poorer provinces in South Africa with a high proportion of rural and unemployed people [22] and people generally migrate from Eastern Cape to the more wealthy western Cape.

Oesophageal cancer is a disease of old age and this applies to women as well and this is supported by the results.

The data collected indicates that the excess risk of oesophageal cancer in patients that attend GSH is associated with smoking. These patients are predominantly elderly persons, born in rural areas, and use of paraffin and wood for cooking. However if one were dissect these associations it looks as though smoking is a stronger risk factor for the coloured patients and living in rural areas. The use of wood is also a strong risk factor and it appears that it is more significant for blacks in this sample.

The gender relationship is a bit distorted here because of the overwhelmingly female component of the controls. This means that the analysis mainly provides information regarding the risk factors that relate to the occurrence of oesophageal cancer in women.

This was confirmed by the results of the analysis for just the women in the sample. This would explain why smoking is a stronger predictor of disease for coloureds in the sample than blacks. Coloured women generally smoke more than black women[22]. This tendency is stronger in coloured of lower socio-economic bracket, hence the rural components. This could relate to level of education of these women although this was not explored in this study. For black women however paraffin and wood used for cooking in rural areas and probably townships are stronger risk factors. It is probably more correct to say that it is not so much the wood and the paraffin but rather the fumes given off when these are burnt that are the likely suspects.

Development of cancer is a multi-step process that occurs over a long period of time, therefore one would be tempted to hypothesise that those who currently use certain fuels were more likely to have used it in the past thus establishing a persistent exposure. This was not found to be true for this samples as there was no association between the use of a certain fuel currently and twenty years ago.

There were three categories for Race: black, coloured and unspecified. One of the inclusion criteria for the study was people of race category Black and coloured, although how this was determined exactly is unclear.

These people then answered a question about what race they were part of. From the model it emerges that people who categorise themselves as coloured have a 53% reduced risk of developing oesophageal cancer than black people in the cancer ward at GSH. People who would not categorise themselves as either black or coloured have an even lower risk, 84% reduced risk, than their black counterparts.

It is very then plausible that these people have distinct characteristics that reduce their risk for developing oesophageal cancer. This, in a way, illustrates the drawbacks of using race groupings to categorise people in South Africa as these population groups have become very heterogeneous in terms of socio economic status, level of education, all of which are important predictors of health.

Most of the data on oesophageal cancer in Africa emphasise that it mainly occurs in men and there is a dearth of information concerning the risk factors that impacts mostly on black and coloured African women. This study brings new insight into the risk of developing oesophageal cancer for women, particularly women in Eastern and Western Cape of South Africa.

There has been previous report that smoking is a risk factor lung cancer in women although it did not extend to oesophageal cancer[3]. Furthermore, that study was conducted in black population in the Gauteng province indicating that smoking is a strong risk factor for oesophageal cancer in men rather than women. The results in our study indicates that when smoking behaviour is present, irrespective of what race or population group one belongs to, the people exposed are at higher risk of oesophageal cancer. The reported variations between women of different race categories in the Western Cape are more indicative of behavioural trends of women of different cultural backgrounds in South Africa. I presume that if Coloured women were exposed to wood smoke and paraffin fumes as their black counterparts instead of smoking they are likely to have a similar risk profile.

Indeed the report about wood-smoke being a risk factor for oesophageal cancer in rural poor women is not the first as a Chinese study also showed that this is a risk factor for women in rural china[18].

What is also interesting is that poor rural coloured women as a whole have lower risk of oesophageal cancer than their black counterparts. Could this be indicative of a stronger link between wood smoke, paraffin fumes with oesophageal cancer or is there another exposure factor that unique to poor black women that results in higher risk. These questions are however not answered in this study. It appears that squamous cell carcinoma is associated with wood-smoke, cigarette smoke, paraffin fumes and low socio economic status but what is still missing is evidence that explains why the OR are so much bigger for wood and paraffin than cigarette smoking. What is needed is a study looking at dose responses in each exposure.

This data however gives indicators that can be used to guide intervention in different areas as risk factors for oesophageal cancer differ across South Africa due to social circumstances.

6. References

1. Hendricks D, Parker MI (2002). *Oesophageal cancer in Africa.* IUBMB *Life* 53: 263-268
2. Enzinger PC, Mayer RJ (2003). *Oesophageal cancer.* The New England Journal Of Medicine 349: 2241-52
3. Pacella-Norman R, Urban MI, Sitas F, Carrara H, Sur R, Hale M, Ruff P, Patel M, Newton R, Bull D, Beral V (2002). *Risk factors for oesophageal, lung oral and laryngeal cancers in black South Africans.* British Journal Of Cancer 86: 1751-1756
4. Segal I, Reinach SG, de Beer M (1988). *Factors associated with oesophageal cancer in Soweto*, South Africa. *Br J Cancer* 58: 681-686.
5. Summeruk R, Segal I, Tewinkel W, Van der Merwe CF (1992). *Oesophageal cancer in three regions of South Africa.* SAMJ 81: 91-93.
6. Rose EF, Fellingham SA (1984). *Cancer patterns in the Transkei.* SAMJ 77: 555-561.
7. Blot WJ (1194). *Oesophageal cancer trends and risk factors.* Seminars in Oncology 21(4): 403-410.
8. Miller JAG, Rege RV, Ko CY, Linvingston EH (2004). *Health care access and poverty do not explain the higher oesophageal cancer mortality in African Americans.* The American Journal of Surgery 188: 22-26.
9. Terry PD, Langergren J, Wolk A, Steineck G, Nyrén O (2003). *Dietary intake of heterocyclic amines and cancers of the oesophagus and gastric cardia.* Cancer Epidemiology, Biomarkers and prevention 12:940-944.
10. Langergren J, Ye W, Lindgren A, Nyrén O (2000). *Heredity and risk of oesophageal cancer of the oesophagus and gastric cardia.* Cancer Epidemiology, Biomarkers and prevention 9: 757-760.

11. White R, Abnet CC, Mungatana CK, Dawsey SM (2002). *Oesophageal cancer: a common malignancy in young people of Bomet district, Kenya.* The Lancet 360: 462-463.

12. McGlashan ND, Harington JS, Cheskowska E (2003). *Changes in the geographical and temporal pattern of cancer incidence among gold miners working in South Africa, 1964-1996.* British Journal of Cancer 88: 1361-1369.

13. Portale G, Peters JH, Hsieh CC, Tamhankar AP, Almogy G, Hagen JA, Demeester SR, Bremner CG, Demeester TR (2001). *Oesophageal Adenocarcinoma in patients ≤ 50 years old: Delayed diagnosis and advanced disease at presentation.* The American surgeon 70(11): 954-958.

14. Gammon MD, Schoenberg JB, Ahsan H, Risch HA, Vaughan T, Chow WH, Rotterdam H, West AB, Dubrow R, Stanford JL, Mayne ST, Farrow DC, Niwa S, Blot WJ, Fraumeni JF Jr. (1997). *Tobacco, Alcohol and socio-economic status and Adenocaricnoma of the oesophagus and gastric cardia.* Journal of the National Cancer Institute 89(17): 1277-1284.

15. Brewster DH, Fraser LA, McKinney PA, Black RJ (2000). *Socio-economic status and risk of adenocaricnoma of the oesophagus and cancer of the gastric cardia in Scotland.* British Journal of Cancer 83(3): 387-390.

16. Surgeon General (1986) Health consequences of using smokeless tobacco. DHHS Washington.

17. Van Rensburg SJ (1981). Epidemiologic and dietary evidence for a specific nutritional predisposition to oesophageal cancer. JNCI 67: 243-251.

18. Cheng KK (1994). *The aetiology of oesophageal cancer in Chinese.* Seminars in Oncology 21(4): 411-415.

19. Spechler ST (1994). Barrett's oesophagus. Seminars in Oncology 21(4): 431-437.

20. Somdyala NIM, Marasa WFO, Venter FS, Vismer HF, gelderblom WCA, Swanevelder SA (2003). *Cancer pattern in four districts of the Transkei region — 1991-1995.* SAMJ 93 (2): 144 – 148.

21. International Classification of diseases, revision 10. www.statsa.go.za www.wolfbane.com/icd/icd10h.htm

22. Eastern Cape province of South Africa. www.ecprov.gov.za

7. Appendix

7.1 Exploratory comparisons

Table 8: Table of exploratory logistic regression models relating risk of cancer to individual risk factors.

Model	Reference	Variable	O R	S.E	P >|z|	95% C. I.
Icdo - age		age	**1.022**	0.008	**0.003**	**1.008 ; 1.038**
Icdo-race	R2 (coloured)	R1 (black)	**1.792**	0.367	**0.004**	**1.199 ; 2.677**
	R2 (coloured)	R3 unspecified	0.337	0.197	0.062	0.108 ; 1.058
Icdo-gender	Female	gender	**18.919**	5.33	**0.000**	**10.669 ; 33.551**
Icdo-birthlocal	Rural	Birthlocal (Urban)	**0.395**	0.077	**0.000**	**0.27 ; 0.578**
Icdo-nowlocal	Rural	Nowlocal (Urban)	0.611	0.159	0.058	0.368 ; 1.016
Icdo-Ecook20	e6 (electricity)	e1 (wood)	**3.14**	0.8	**0.000**	**1.91 ; 5.159**
		e3 (coal)	1.29	0.64	0.608	0.486 ; 3.428
		e4 (paraffin)	**2.25**	0.55	**0.001**	**1.387 ; 3.428**
		e5 (gas)	1.5	0.62	0.328	0.666 ; 3.375
		e7 (other)	3.23	3.98	0.342	0.287 ; 36.268
Icdo-Ecooknow	E1(electricity)	E2 (gas)	0.842	0.329	0.661	0.392 ; 1.812
		E3 (paraffin)	**2.663**	0.687	**0.000**	**1.606 ; 4.414**
		E4 (wood)	**2.819**	0.998	**0.003**	**1.408 ; 5.641**
Icdo- smoke	S3(never)	s1 (current)	**4.561**	1.066	**0.000**	**2.885 ; 7.212**
		s2 (in past)	**3.801**	0.957	**0.000**	**2.321 ; 6.225**

- e2 was dropped from the model due to co-linearity.
- E5 was dropped from the model because there is only 1 observation.
- S.E. is the standard error for each variable.
- O.R. is the Odds Ratio.

Box 1: Logistic regression of current energy source for cooking on gender

```
. logistic gender E2 E3 E4 E5

note: E5~=0 predicts failure perfectly
      E5 dropped and 1 obs not used


Logit estimates                                   Number of obs   =        451
                                                  LR chi2(3)      =      17.13
                                                  Prob > chi2     =     0.0007
Log likelihood = -279.64473                       Pseudo R2       =     0.0297

------------------------------------------------------------------------------
      gender | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          E2 |   .5185185   .2324846    -1.46   0.143     .2153361    1.248566
          E3 |   1.041152   .2623943     0.16   0.873     .6353191    1.706226
          E4 |   .1793372    .096744    -3.19   0.001     .0622999    .5162422
```

Box 2: logistic regression of current energy source for cooking and gender on risk of cancer (icdo)

```
logistic icdo gender E2 E3 E4 E5

note: E5~=0 predicts failure perfectly
      E5 dropped and 1 obs not used


Logit estimates                                   Number of obs   =        451
                                                  LR chi2(4)      =     197.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -213.73703                       Pseudo R2       =     0.3162

------------------------------------------------------------------------------
        icdo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   28.17349   8.828006    10.65   0.000     15.24479    52.06668
          E2 |   1.391076   .6698549     0.69   0.493     .5413345    3.574671
          E3 |   4.276934   1.340409     4.64   0.000     2.313995    7.905016
          E4 |   8.220153   3.176237     5.45   0.000     3.854595    17.52997
```

Box 3: logistic regression of current energy source for cooking and gender on risk of cancer (icdo) including the interaction term cookgender

```
. logistic icdo gender E2 E3 E4 E5  cookgender

note: E5~=0 predicts failure perfectly
      E5 dropped and 1 obs not used


Logit estimates                                   Number of obs   =        451
                                                  LR chi2(5)      =     197.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -213.65985                       Pseudo R2       =     0.3164

------------------------------------------------------------------------------
        icdo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   35.31671   22.92217     5.49   0.000     9.897172    126.0229
          E2 |   1.435571   .6954776     0.75   0.455     .5554578    3.710208
          E3 |   4.452978   1.463639     4.54   0.000     2.338143    8.480666
          E4 |   8.390339   3.270786     5.46   0.000     3.908028    18.01364
  cookgender |   .8404764   .3595703    -0.41   0.685     .3633849    1.943946
------------------------------------------------------------------------------
```

Box 4: regression of s1 and s2 on age

```
. Regress age s1 s2

      Source |       SS       df       MS              Number of obs =     452
-------------+------------------------------           F(  2,   449) =    7.58
       Model |  2320.68189     2  1160.34095            Prob > F      =  0.0006
    Residual |  68752.2628   449   153.12308            R-squared     =  0.0327
-------------+------------------------------           Adj R-squared =  0.0283
       Total |  71072.9447   451  157.589678            Root MSE      =  12.374

------------------------------------------------------------------------------
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          s1 |  -.6506239   1.348942    -0.48   0.630    -3.301647      2.0004
          s2 |   4.915202   1.478516     3.32   0.001     2.009532    7.820873
       _cons |   56.51337   .9048978    62.45   0.000     54.73501    58.29173
```

Box 5: logistic model of energy source for cooking on birthlocal

```
logistic  birthlocal   E2 E3 E4 E5

note: E5~=0 predicts success perfectly
      E5 dropped and 1 obs not used


Logit estimates                                   Number of obs   =        451
                                                  LR chi2(3)      =      97.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -262.72427                       Pseudo R2       =     0.1567

------------------------------------------------------------------------------
  birthlocal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          E2 |   .5340991   .2066531    -1.62   0.105     .2501933    1.140166
          E3 |    .178033   .0482064    -6.37   0.000     .1047174    .3026789
          E4 |   .0233668   .0171834    -5.11   0.000      .005529    .0987535
------------------------------------------------------------------------------
```

Box 6: logistic model of race on birthlocal

```
logistic  birthlocal  R2 R3

Logit estimates                                   Number of obs   =        452
                                                  LR chi2(2)      =     132.16
                                                  Prob > chi2     =     0.0000
Log likelihood = -246.09116                       Pseudo R2       =     0.2117

------------------------------------------------------------------------------
  birthlocal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          R2 |   11.84325   3.160815     9.26   0.000     7.019337    19.98232
          R3 |   37.20833   38.45921     3.50   0.000     4.907077    282.1354
------------------------------------------------------------------------------
```

Box 7: mlogit model of energynow on energy20

```
. mlogit  energy20 energynow, rrr base(6)

Iteration 0:   log likelihood =  -635.9422
Iteration 1:   log likelihood = -635.46312
Iteration 2:   log likelihood = -635.42275
Iteration 3:   log likelihood = -635.38954
Iteration 4:   log likelihood =  -634.9043
Iteration 5:   log likelihood = -634.46245
Iteration 6:   log likelihood =  -634.4448
Iteration 7:   log likelihood = -634.44479

Multinomial logistic regression                 Number of obs   =        452
                                                LR chi2(6)      =       2.99
                                                Prob > chi2     =     0.8095
Log likelihood = -634.44479                     Pseudo R2       =     0.0024

------------------------------------------------------------------------------
    energy20 |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1 WOOD       |
   energynow |   .9983343   .0033707    -0.49   0.621     .9917496    1.004963
-------------+----------------------------------------------------------------
2 CHARCOAL   |
   energynow |   .7765577   .5971689    -0.33   0.742       .17203    3.505447
-------------+----------------------------------------------------------------
3 COAL       |
   energynow |   .7249204   .1981207    -1.18   0.239     .4242843    1.238579
-------------+----------------------------------------------------------------
5 PARAFIN    |
   energynow |   .9973046   .0054134    -0.50   0.619     .9867507    1.007971
-------------+----------------------------------------------------------------
6 GAS        |
   energynow |   .8920765   .1748194    -0.58   0.560     .6075653    1.309819
-------------+----------------------------------------------------------------
999          |
   energynow |   .9242841   .5184853    -0.14   0.888     .3078339    2.775201
------------------------------------------------------------------------------
(Outcome energy20==7 ELECTRICITY is the comparison group)
```

7.2 Model building

Model A

```
. logit icdo

Iteration 0:   log likelihood =  -313.2627

Logit estimates                                 Number of obs   =        452
                                                LR chi2(0)      =      -0.00
                                                Prob > chi2     =          .
Log likelihood =  -313.2627                     Pseudo R2       =    -0.0000

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .0265502   .0940804     0.28   0.778    -.1578439    .2109444
------------------------------------------------------------------------------
```

Model B

```
. logit icdo birthlocal

Iteration 0:   log likelihood =  -313.2627
Iteration 1:   log likelihood = -301.49318
Iteration 2:   log likelihood = -301.48901

Logit estimates                                   Number of obs   =        452
                                                  LR chi2(1)      =      23.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -301.48901                       Pseudo R2       =     0.0376

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  birthlocal |  -.9281159   .1939124    -4.79   0.000    -1.308177   -.5480546
       _cons |   .5260931   .1428152     3.68   0.000     .2461805    .8060057
------------------------------------------------------------------------------
```

Model C

```
. logit icdo gender

Iteration 0:   log likelihood =  -313.2627
Iteration 1:   log likelihood = -239.53734
Iteration 2:   log likelihood = -236.94749
Iteration 3:   log likelihood = -236.87765
Iteration 4:   log likelihood = -236.87755

Logit estimates                                   Number of obs   =        452
                                                  LR chi2(1)      =     152.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -236.87755                       Pseudo R2       =     0.2438

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   2.940185   .2922948    10.06   0.000     2.367298    3.513073
       _cons |  -.8001193   .1248343    -6.41   0.000     -1.04479   -.5554485
------------------------------------------------------------------------------
```

Model D

```
. logit icdo gender birthlocal

Iteration 0:   log likelihood =  -313.2627
Iteration 1:   log likelihood = -223.33387
Iteration 2:   log likelihood = -217.59158
Iteration 3:   log likelihood = -217.35835
Iteration 4:   log likelihood = -217.35772

Logit estimates                                   Number of obs   =        452
                                                  LR chi2(2)      =     191.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -217.35772                       Pseudo R2       =     0.3061

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    3.28792   .3186951    10.32   0.000     2.663288    3.912551
  birthlocal |  -1.493119    .252742    -5.91   0.000    -1.988484   -.9977533
       _cons |   -.119961   .1637024    -0.73   0.464    -.4408119    .2008899
------------------------------------------------------------------------------
```

Model E

```
. logit icdo gender birthlocal age

Iteration 0:   log likelihood =  -313.2627
Iteration 1:   log likelihood =  -221.3595
Iteration 2:   log likelihood = -215.40292
Iteration 3:   log likelihood = -215.15807
Iteration 4:   log likelihood = -215.15738

Logit estimates                                   Number of obs   =        452
                                                  LR chi2(3)      =     196.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -215.15738                       Pseudo R2       =     0.3132

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    3.266045   .3188856    10.24   0.000     2.641041     3.89105
  birthlocal |   -1.468944   .2537103    -5.79   0.000    -1.966207    -.9716805
         age |    .0201082   .0096875     2.08   0.038     .0011212     .0390953
       _cons |   -1.291867   .5894975    -2.19   0.028    -2.447261    -.136473
------------------------------------------------------------------------------
```

Model F

```
. logit icdo gender birthlocal age  E2 E3 E4 E5

note: E5 != 0 predicts failure perfectly
      E5 dropped and 1 obs not used

Iteration 0:   log likelihood = -312.55505
Iteration 1:   log likelihood = -212.24581
Iteration 2:   log likelihood =  -204.9443
Iteration 3:   log likelihood = -204.58324
Iteration 4:   log likelihood = -204.58184

Logit estimates                                   Number of obs   =        451
                                                  LR chi2(6)      =     215.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -204.58184                       Pseudo R2       =     0.3455

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    3.454848   .3282929    10.52   0.000     2.811406    4.098291
  birthlocal |   -1.019099   .2779726    -3.67   0.000    -1.563915   -.4742823
         age |    .0200066   .0101341     1.97   0.048     .0001442    .0398691
          E2 |    .0961767   .4995561     0.19   0.847    -.8829353    1.075289
          E3 |    1.224414   .3337455     3.67   0.000     .5702849    1.878543
          E4 |     1.42674   .4166611     3.42   0.001     .6100991    2.243381
       _cons |   -1.964774   .6352032    -3.09   0.002     -3.20975   -.7197987
------------------------------------------------------------------------------
```

Model G

```
. logit icdo gender birthlocal age  E2 E3 E4 E5 R2 R3

note: E5 != 0 predicts failure perfectly
      E5 dropped and 1 obs not used

Iteration 0:   log likelihood = -312.55505
Iteration 1:   log likelihood = -210.79664
Iteration 2:   log likelihood = -202.89183
Iteration 3:   log likelihood = -202.40735
Iteration 4:   log likelihood = -202.40396
Iteration 5:   log likelihood = -202.40396

Logit estimates                                   Number of obs   =        451
                                                  LR chi2(8)      =     220.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -202.40396                       Pseudo R2       =     0.3524

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    3.499444   .3367074    10.39   0.000      2.83951    4.159378
  birthlocal |   -.8442861   .2984034    -2.83   0.005    -1.429146   -.2594261
         age |    .0223486   .0102589     2.18   0.029     .0022415    .0424557
          E2 |   -.0138362    .508922    -0.03   0.978    -1.011305    .9836327
          E3 |    1.064765   .3507905     3.04   0.002     .3772286    1.752302
          E4 |    1.311624   .4244106     3.09   0.002     .4797947    2.143454
          R2 |   -.3526137   .3337814    -1.06   0.291    -1.006813    .3015858
          R3 |   -1.578483   .8258028    -1.91   0.056    -3.197027    .0400603
       _cons |    -1.98716   .6398373    -3.11   0.002    -3.241218   -.7331021
------------------------------------------------------------------------------
```

Model H

```
. logit icdo gender birthlocal age  E2 E3 E4 E5 s1 s2

note: E5 != 0 predicts failure perfectly
      E5 dropped and 1 obs not used

Iteration 0:   log likelihood = -312.55505
Iteration 1:   log likelihood = -208.07565
Iteration 2:   log likelihood = -199.61038
Iteration 3:   log likelihood = -199.09286
Iteration 4:   log likelihood =  -199.0897
Iteration 5:   log likelihood =  -199.0897

Logit estimates                                   Number of obs   =        451
                                                  LR chi2(8)      =     226.93
                                                  Prob > chi2     =     0.0000
Log likelihood =  -199.0897                       Pseudo R2       =     0.3630

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    3.216003   .3521677     9.13   0.000     2.525767    3.906239
  birthlocal |   -1.234879   .2938681    -4.20   0.000    -1.810849   -.6589077
         age |    .0238532   .0105376     2.26   0.024     .0031999    .0445065
          E2 |    .0226789   .5030081     0.05   0.964    -.9631988    1.008557
          E3 |    1.203726   .3390146     3.55   0.000     .5392694    1.868182
          E4 |    1.384909   .4279264     3.24   0.001     .5461888    2.223629
          s1 |    1.015569   .3223709     3.15   0.002     .3837334    1.647404
          s2 |    .1526216    .336138     0.45   0.650    -.5061968    .8114401
------------------------------------------------------------------------------
```

_cons	-2.359309	.66826	-3.53	0.000	-3.669074	-1.049543

Model I

```
. logit icdo gender birthlocal age  E2 E3 E4 E5 s1 s2 R2 R3

note: E5 != 0 predicts failure perfectly
      E5 dropped and 1 obs not used

Iteration 0:   log likelihood = -312.55505
Iteration 1:   log likelihood = -205.51811
Iteration 2:   log likelihood = -196.00949
Iteration 3:   log likelihood = -195.27212
Iteration 4:   log likelihood = -195.26422
Iteration 5:   log likelihood = -195.26422

Logit estimates                                   Number of obs   =        451
                                                  LR chi2(10)     =     234.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -195.26422                       Pseudo R2       =     0.3753

------------------------------------------------------------------------------
        icdo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   3.230542   .3628158     8.90   0.000     2.519436    3.941648
  birthlocal |  -.9728576   .3096276    -3.14   0.002    -1.579717   -.3659988
         age |   .0275859   .0107616     2.56   0.010     .0064935    .0486783
          E2 |   -.181548   .5167463    -0.35   0.725    -1.194352    .8312561
          E3 |   .9116125   .3589377     2.54   0.011     .2081076    1.615118
          E4 |   1.160299   .4389217     2.64   0.008     .3000286     2.02057
          s1 |   1.245603   .3467153     3.59   0.000     .5660532    1.925152
          s2 |   .2446966   .3420705     0.72   0.474    -.4257493    .9151425
          R2 |  -.7351193   .3609615    -2.04   0.042    -1.442591   -.0276477
          R3 |  -1.819139   .8112221    -2.24   0.025    -3.409105   -.2291732
       _cons |  -2.409953   .6758884    -3.57   0.000     -3.73467   -1.085236
```

Table 9: Summary of model building and selection

Risk factors in Model	LL	LR Chi2	P	AIC
A: constant	-312.26			628.53
B: birthlocal	-301.49	23.55	0.000	606.98
C: gender	-236.88	152.77	0.000	477.76
D: gender birthlocal	-217.36	39.4	0.000	440.72
E: gender birthlocal age	-215.16	4.4	0.0359	438.31
F: gender birthlocal age E2 E3 E4	-204.58			423.16
G: gender birthlocal age E2 E3 E4 R2 R3	-202.40	4.36	0.1133	422.81
H: gender birthlocal age E2 E3 E4 s1 s2	-199.09	10.97	0.0041	416.17
I: gender birthlocal age E2 E3 E4 s1 s2 R2 R3	**-195.26**	**7.65**	**0.0218**	**412.53**

LL: Log likelihood
LR Chi2: likelihood ratio test
P: Prob>chi2
AIC: Aikaike's information criterion
Note: Lrtest for model F nested in G was not significant, p = 0.113

Final model for the whole sample

```
. logistic icdo gender birthlocal age E2 E3 E4 E5 s1 s2 R2 R3
note: E5~=0 predicts failure perfectly
      E5 dropped and 1 obs not used

Logit estimates                                   Number of obs   =        451
                                                  LR chi2(10)     =     234.58
                                                  Prob > chi2     =      0.000
Log likelihood = -195.26422                       Pseudo R2       =      03753
------------------------------------------------------------------------------
        icdo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Gender |   25.29337   9.176833     8.90   0.000     12.42159    51.50342
 birth local |   .3780013   .1170396    -3.14   0.002     .2060335    .6935037
         Age |    1.02797   .0110626     2.56   0.010     1.006515    1.049883
       Gas C |   .8339782   .4309552    -0.35   0.725     .3029001    2.296201
   Paraffin C |   2.488332   .8931561     2.54   0.011     1.231346    5.028479
      Wood C |   3.190888    1.40055     2.64   0.008     1.349897    7.542623
    C smokers |   3.475029   1.204846     3.59   0.000     1.761302    6.856193
Past smokers |   1.277234    .436904     0.72   0.474     .6532801    2.497131
   Coloureds |   .4794482   .1730624    -2.04   0.042     .2363147     .972731
        Race |   .1621653   .1315521    -2.24   0.025     .0330708    .7951908
(Unspecified)
C: Current
```

Final Model with women only

```
. logistic icdo  birthlocal age E2 E3 E4 E5 s1 s2 R2 R3 if gender==0

note: E5 != 0 predicts failure perfectly
      E5 dropped and 1 obs not used


Logistic regression                               Number of obs   =        299
                                                  LR chi2(9)      =      76.02
                                                  Prob > chi2     =     0.0000
Log likelihood = -147.34932                       Pseudo R2       =     0.2051

------------------------------------------------------------------------------
        icdo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  birthlocal |   .3437463   .1167016    -3.15   0.002     .1767079    .6686829
         age |   1.034587   .0126171     2.79   0.005     1.010151    1.059614
          E2 |    .670513    .389772    -0.69   0.492     .2145854    2.095145
          E3 |   2.799699   1.070075     2.69   0.007     1.323649    5.921746
          E4 |   2.900703   1.314501     2.35   0.019     1.193348    7.050814
          s1 |   3.240709   1.282953     2.97   0.003      1.49163    7.040754
          s2 |   1.466623   .5636104     1.00   0.319     .6905721    3.114783
          R2 |   .4586156   .1968896    -1.82   0.069     .1977047     1.06385
          R3 |   .2407895   .2656746    -1.29   0.197     .0276991    2.093193
------------------------------------------------------------------------------
```

7.3 Model checking

Table 10: potential outliers and influential observations

	icdo	p	h	d2	rs	ds
26.	**0**	**.9762192**	**.0054065**	**41.27391**	**-6.424477**	**-2.741971**
40.	**0**	**.9626524**	**.0089288**	**26.00775**	**-5.09978**	**-2.575696**
45.	**0**	**.9793521**	**.0050581**	**47.67207**	**-6.904496**	**-2.792799**
70.	0	.912488	.0240728	4.36657	-2.089634	-1.529324
81.	1	.89579	.0487718	7.056498	-2.656407	-2.035299
87.	0	.8313909	.0259067	5.062016	-2.249892	-1.911814
91.	0	.8384767	.0252283	5.325408	-2.307685	-1.934058
119.	0	.8974479	.0202389	8.931913	-2.988631	-2.156121
147.	0	.670184	.0702214	4.37092	-2.090674	-2.184485
155.	0	.6693768	.0726465	4.366385	-2.08959	-2.184928
203.	1	.356799	.0370452	3.744097	1.934967	2.069038
215.	1	.2024364	.0501394	4.147792	2.036613	1.833928
264.	1	.1110059	.0617791	8.535865	2.92162	2.164678
296.	1	.2224673	.0348513	7.242496	2.691189	2.495795
315.	1	.2224673	.0348513	7.242496	2.691189	2.495795
328.	1	.1118148	.0185085	8.093151	2.844846	2.112925
333.	1	.1382354	.0408823	6.499762	2.549463	2.031326
346.	1	.0825946	.0191528	4.688476	2.165289	1.560149
364.	1	.1131774	.0113924	7.925979	2.815311	2.099479
374.	**1**	**.0883019**	**.0118387**	**10.44848**	**3.23241**	**2.216335**
391.	0	.89579	.0487718	7.056498	-2.656407	-2.035299
392.	1	.356799	.0370452	3.744097	1.934967	2.069038
394.	0	.670184	.0702214	4.37092	-2.090674	-2.184485
397.	0	.89579	.0487718	7.056498	-2.656407	-2.035299
409.	0	.6693768	.0726465	4.366385	-2.08959	-2.184928
413.	1	.1384407	.0218999	6.362648	2.522429	2.010764
424.	1	.1113946	.0145089	8.094539	2.84509	2.110444
431.	1	.0936657	.0110991	9.784864	3.128077	2.188424
434.	0	.0825946	.0191528	4.688476	2.165289	1.560149
441.	1	.89579	.0487718	7.056498	-2.656407	-2.035299
442.	1	.912488	.0240728	4.36657	-2.089634	-1.529324
452.	0	.8371587	.0236593	5.265529	-2.294674	-1.928191

Table 10 contains a list of potential outliers and influential observations in the model with the whole sample. Other than observations 26, 40, 45 and 374, all other observation had residuals ordering around 2 or –2, d2 lower than 10 and h values close to zero. Observation 26, 40, 45 and 374 had odd probability values for their actual diagnoses. 26, 40 and 45 according to

their risk factor profiles had very high probabilities for developing oesophageal cancer but were negative and 374 had a very low probability but had a positive diagnosis.

Table 11: potential influential observations from the model with just women

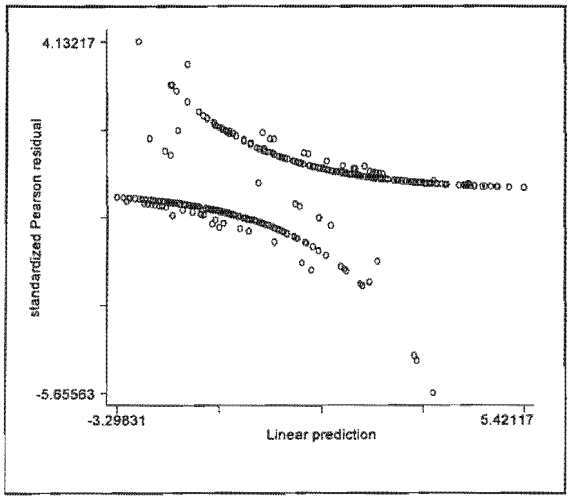
	icdo	p	h	d2	rs	ds
364.	1	0.1009148	0.0124566	9.021726	3.003619	2.155183
431.	1	0.0894311	0.0141823	10.32827	3.213763	2.213152
328.	1	0.0804023	0.0187838	11.65641	3.414149	2.266701
374.	1	0.0739785	0.012516	12.6761	3.560351	2.296513

Box 8: **Model with women only and potentially influential observations removed**

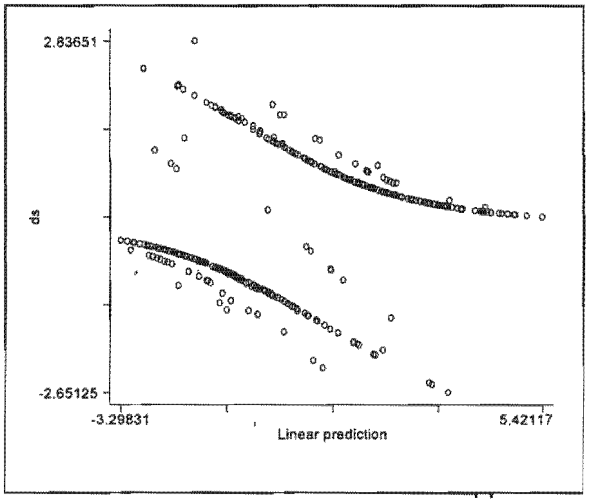
```
. logistic icdo  birthlocal age E2 E3 E4 E5 s1 s2 R2 R3 if gender==0

note: E5 != 0 predicts failure perfectly
      E5 dropped and 1 obs not used


Logistic regression                               Number of obs   =        295
                                                  LR chi2(9)      =      87.75
                                                  Prob > chi2     =     0.0000
Log likelihood = -136.75342                       Pseudo R2       =     0.2429

------------------------------------------------------------------------------
        icdo | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  birthlocal |    .270904   .0975637    -3.63   0.000     .1337404    .5487422
         age |   1.043424   .0136846     3.24   0.001     1.016945    1.070593
          E2 |   .7193684   .4286948    -0.55   0.580     .2237139    2.313182
          E3 |   3.280729   1.301417     2.99   0.003     1.507687     7.13887
          E4 |   3.006623   1.388037     2.38   0.017     1.216504    7.430952
          s1 |   3.766953   1.581709     3.16   0.002     1.654164    8.578312
          s2 |   1.607338   .6342487     1.20   0.229      .741697    3.483274
          R2 |   .4300176   .1966319    -1.85   0.065     .1754937    1.053685
          R3 |   .2984188   .3326308    -1.08   0.278     .0335768    2.652242
------------------------------------------------------------------------------
```

Graphs 2 to 6 are graphs of the various diagnostic indicators mentioned above. These indicate that the model is correct.

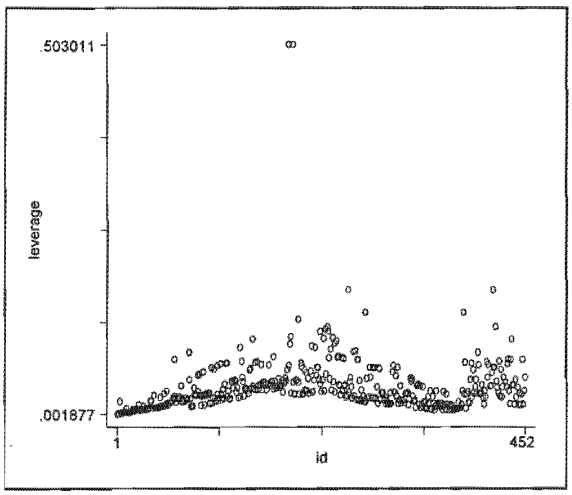
Graph 2: graph of rs against linear predictor (xb)



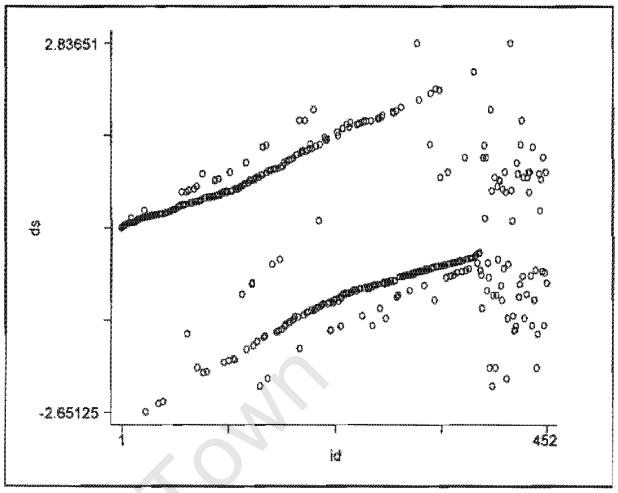
Graph 3: graph of ds against linear predictor (xb)



Graph 4: graph of h against observation(id)



Graph 5: graph of ds against observation(id)



Graph 6: graph of d2 against observation(id)