University of Sheffield
Department of Probability and Statistics

Lung Cancer Prediction

Zahra [redacted]sari

This dissertation is submitted in partial fulfilment of the
requirements for the
MSc in Statistics with Medical Applications

September 2011

Abstract of Lung Cancer Risk Prediction

Zahra [redacted]sari

September 2011

BACKGROUND: This project involves applying up to three lung cancer risk prediction models to the database of 516 patients and 363 controls comprising the ReSoLuCENT (Resource for the study of lung cancer in North Trent) series.

METHODS: The Bach [1] and the Liverpool Lung project (LLP) [2] models have been developed recently to estimate individual's absolute lung cancer risk. The two models share risk factors such as smoking duration and occupational exposure to asbestos. Models vary in the inclusion of lung related co morbidities or family history. The discriminatory power, accuracy, PPV (positive predictive value) and NPV (negative predictive value) of these models have been compared for 5-year lung cancer risk. This is in a manner similar to D'Amelio Jr, AM; Cassidy, Asomaning, K, et al. (2010) Comparison of discriminatory power and accuracy of three lung cancer risk models; British Journal of cancer [3]. Model's discriminatory power has been compared by calculating the area under the curve (AUC) of the receiver operator characteristic (ROC) curve. Data handling and analysis have been conducted using R.

RESULTS: Overall, the Liverpool Lung Project (LLP) model had a high discriminatory power 0.72, whereas the Bach model had significantly lower power 0.60. Positive predictive values were slightly higher with Bach model but negative predictive values were higher for the LLP models. The LLP model had lower sensitivity but better specificity than did the Bach model.

CONCLUSION: Moderate differences were observed in discriminatory power among the lung cancer risk models confirming the difficulty in developing effective risk models. LLP model was favourable over Bach model for higher discriminatory power and inclusion of family history.

Contents

1 Introduction

One in four (27%) of all deaths in the UK are caused by cancer. There were 156,723 cancer deaths in the UK in 2008. With 35,261 lung cancer deaths in 2008, lung cancer is the most common cause of death from cancer for both men and women, causing 24% of all male cancer deaths and 21% of all female cancer deaths.

The lifetime risk of developing lung cancer in 2008 has been estimated to be 1 in 14 for men and 1 in 19 for women in the UK. Overall more than 1 in 5 (22%) of all cancer deaths are from lung cancer. Lung cancer was the most frequently occurring cancer in the UK; until in 1997 it was overtaken by breast cancer but still accounts for around 1 in 8 new cancer cases, that is, 40,806 new cases diagnosed in 2008.

Cigarette smoking is the single most important cause of preventable death in the UK. Smokers are, on average, much more likely to get cancer than non-smokers. Around half of current smokers will be killed by their habit if they continue to smoke. More than a quarter of all deaths from cancer, including an estimated 90% of lung cancer deaths, are linked to tobacco smoking. Studies from Europe, Japan and North America have shown that 9 in 10 lung cancers are caused by smoking [7]. Lung cancer risk is greatest among those who start smoking young and smoke the most cigarettes over the longest period of time. The length of time spent smoking seems to be more important than the other two factors. The risk of developing lung cancer among people who have smoked varies widely. Determining individual risk from an individual's history of exposures would be a useful tool for both patient care and clinical research.

Absolute Risk Assessment Models estimate the probability of developing cancer over a defined period of time (e.g. 5 years)[4]. Accurate cancer risk assessment helps identify and control the suffering and death of average and high risk individuals. Cancer risk prediction models further facilitate the design and planning of clinical chemoprevention trials, enable the development of benefit-risk indices and provide estimates of the population burden and cost of cancer[5].

Lung cancer mortality trends vary considerably over time between genders. For men, age standardised lung cancer mortality rates decreased steadily between 1982 and 2008. Over the same time period, female mortality rates for lung cancer increased until the mid 1990s, then leveled off until 2004, but since then there is a suggestion that rates

may be increasing further. It is predicted that male deaths from lung cancer will continue to fall while female deaths will slightly increase.

Studying the histology of lung cancer, there are two main types : around 20% are small cell lung cancers (SCLC) and the remainder are non-small cell lung cancers (NSCLC). The main types of NSCLC are squamous cell carcinoma, adenocarcinoma and large cell carcinoma. Smoking has been linked to all four types of lung cancer, although adenocarcinoma is the most common type in non-smokers and a rise in incidence has been reported in the USA and other countries. In the USA, adenocarcinoma is now the most common type of lung cancer. In Europe the most common type of lung cancer is still squamous cell carcinoma despite increases in the incidence of adenocarcinoma. The increasing incidence of adenocarcinoma has been linked to low-tar cigarettes. The lung cancer subtypes are known, but not relevant for this project and not studied further.

2 Introduction to ReSoLuCENT Study

Resource for the Study of Lung Cancer Epidemiology in North Trent (ReSoLuCENT) is a study in the NIHR CRN (National Institute for Health Research Clinical Research Network) Portfolio. [8]

This study aims to collect good quality epidemiological and biological data from lung cancer patients with a family history of the disease or with early onset lung cancer. This includes detailed questionnaires to elicit the family history, smoking history and occupational risk factors for lung cancer. Blood samples have been taken for serum proteins, genomic and plasma DNA. Permission is sought to access tumour biopsy samples. Patient cases and controls (their partners and first degree relatives) have been invited to participate. This resource has been available for detailed studies of inherited and acquired genetic changes, and proteomic analysis in lung cancer.

Large collaborative studies play an important role in genetic epidemiology as the impact of genetic variants in different geographical areas and ethnicities can be examined. In a case control study, controls provide the baseline reference group for each population. Cases and controls have been recruited in approximately equal numbers. ReSoLuCENT has a matched design where controls are usually recruited through the case. To get the maximum benefit of the resource, each case is matched with at least one control (ideally one related and one non-related). In the current data set 516 (59%) of controls were the case's brother or sister, 97 (11%) were the case's parent and 51 (6%) were the case's child.

2.1 Recruitment Pathway

Eligible patients have been identified in Lung MDT (Multidisciplinary Team) or clinic with the consultant's agreement. Below is the eligibility criteria for cases and controls.

Cases have lung cancer or need to have an operation for suspected lung cancer who

- Are 60 years old or less or
- Have a first degree relative aged 60 or less who has lung cancer or
- Have two or more first or second degree relatives of any age with lung cancer

Controls are cancer free

- Recruited through Cases, are the partner of someone taking part in the study or
- Are a first degree relative of someone taking part in the study, and are at least 18 years old

Patients are given the Patient Information Sheet and Lifestyle Questionnaire while controls only complete the Lifestyle Questionnaire.(section 11.3)

- If no, refusal is logged
- If yes, the nurse arranges consent, blood sampling and collects and checks the completed Lifestyle Questionnaire.

2.2 Recruitment Sites

Currently ReSoLuCENT study is based in : Airedale, Cardiff, Manchester, Sheffield and Southampton. More sites will be taking part soon. This study will recruit about 2,500 people altogether. This will be about 1,000 cases and 1,500 controls. Any site will need to be able to enter at least 20 - 30 patients per year with the expectation of at least one control per patient (at least 80% of cases must have a control).

- Sheffield and Doncaster
- Southampton General Hospital
- Airedale General Hospital
- Christie Hospital, Manchester
- Velindre Hospital, Cardiff
- Prince Charles Hospital, Merthyr Tidfil
- Royal Glamorgan Hospital, Llantrisant

3 Data Processing

3.1 Crude and Processed Data

ReSoLuCENT main data sheet comprises of 168 variables for 879 subjects. 516 patients and 363 controls have been recruited for the purpose of ReSoLuCENT series from 27/04/2006 to 16/03/2011. 23 raw variables have been used for the purpose of data analysis as in Table 1 .

FieldName	Field Type
CaseControl_c	Case or Control
AgeRegistered	Age at registration or study
Sex	Number (1 =male, 2 =female)
Status	Status (Alive, Dead)
DeathDate	day/ month/ year
DtBirth	day/month/ year
DtDiag	day/month/ year
Sitelung	Affected lung (Left, Right)
RelCase	Relationship to case
S16	Occupation (Self reported free text)
S19	Dusty (Yes , No)
S20	Asbestos Exposure
S28	Regular Smoking (Current, Ex-smoker, Non-smoker)
S29	Age at first full cigarette
S30	Age smoked regularly
S31	Age last smoked , if stopped smoking
S31_20	Cigarettes smoked per day at age 20
S31_30	Cigarettes smoked per day at age 30
S31_40	Cigarettes smoked per day at age 40
S31_50	Cigarettes smoked per day at age 50
S37	Passive smoking
S04_1	Subjects living with children or stepchildren
S04_4	Subjects living with Siblings

Table 1: Main Sheet Raw Data

Pack year has been calculated using a complex algorithm in section 5.

Smoking cessation duration is calculated in R for ex-smokers only by subtracting age registered by age stopped smoking regularly (S31). This is further described in EDA section 3.2 along figure 8 .

Smoking duration is calculated for ex-smokers and current smokers, cases seem to have smoked for longer . Figure 9 is the frequency histogram of smoking duration for cases and controls.

CPD is the weighted average of cigarettes smoked per day, calculated for the Bach risk model in section 6.2.

FieldName	Field Type
Pack Year (p)	Variables: S30, S31, AgeRegistered, S31_20, S31_30, S31_40, S31_50 =(number of cigarettes smoked per day x number of years smoked)/20
Smoking Cessation Duration	= AgeRegistered -S31 = (Age registered - Age stopped smoking regularly)
Smoking Duration	= S31-S30 For ex-smokers = (Age stopped - Age started smoking regularly) For Current Smokers =(AgeRegistered-Age started smoking regularly)
Cigarettes Per Day (CPD)	Current smoker : 20*p/(AgeRegistered-S30) = 20*pack year/ (Age registered - Age started smoking regularly) Ex- smoker : 20*p/(S31-S30) = 20*pack year/(Age stopped -Age started smoking regularly)

Table 2: Processed Data Main Sheet

ReSoLuCENT Health Sheet (data relevant to questions "F", "M", "B" and "Si" of Questionnaire (Section 11.3) comprises of 47 variables recorded for 6443 individuals and includes the family history of smoking (Parents and siblings), cancer and health related information. Table 3 is list of crude data that has been used for the purpose of family history data analysis in section 3.3.

FieldName	Field Type
H1Alive	Family alive
H1Cause	Cause of death
H3	Family smoking status
HCa	Has had cancer
H-2Ca	More than 1 cancer
HCaSite1	1st Part of body affected by cancer
HCaSite2	Cancer Site 2

Table 3: Health Sheet Raw Data

3.2 Exploratory Data Analysis

Figure 1 portrays the relative proportion of cases and controls. The mean age of patients is 55.26 with standard deviation 6.91 while the mean age for controls was 51.27 with standard deviation 11.03. Cases and controls have different distributions due to the fact that lung cancer is mostly diagnosed between ages 50-60 while controls are from a diverse age range.

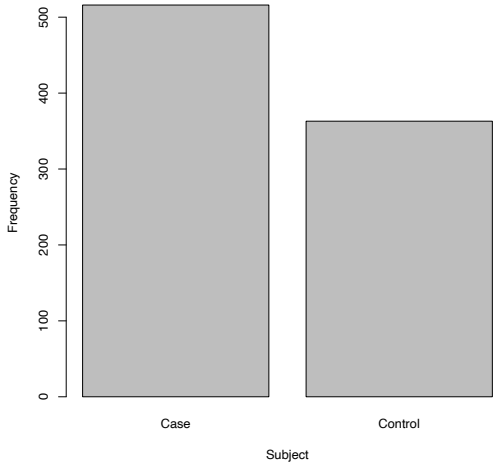

Figure 1: Case - Control Frequency

More than three-quarters of lung cancer sufferers die aged 65 and over. 10% of lung cancer is diagnosed under age 60 but in ReSoLuCENT study due to eligibility criteria older cases are not registered. Total mean age of ReSoLuCENT data is 52.84. Figure 2 is a slightly skewed unimodal distribution of age with peak of 52.84.

Figure 3 is segmented and side-by-side bar plot of gender distribution for cases and controls. The total number of males recruited is 357 while the total number of females is 451. There are 71 missing values where the subject's gender is unknown. The number of cases recruited are equivalent for both genders, 224 for male and 225 for females. The number of controls recruited are 133 for male and 226 for female. Controls were enrolled through cases, ie. they are first-degree relative or a co-habiting partner of a study patient. A large number of patients who lived with their spouse volunteered as control. This explains the negative correlation of gender in cases and controls.

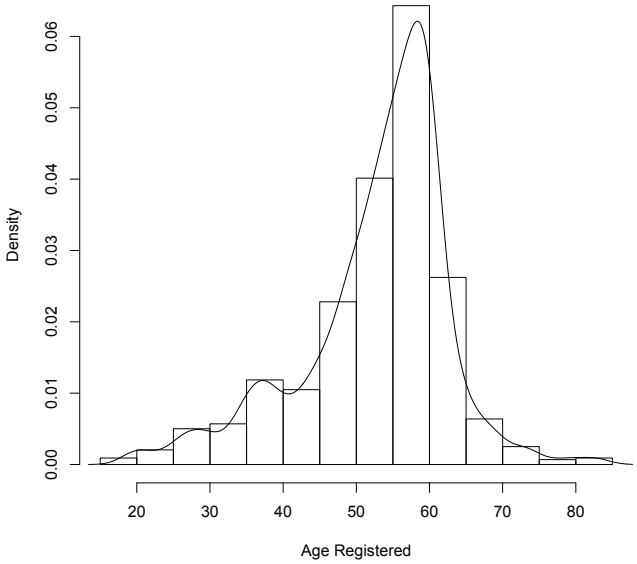

Figure 2: Age Histogram

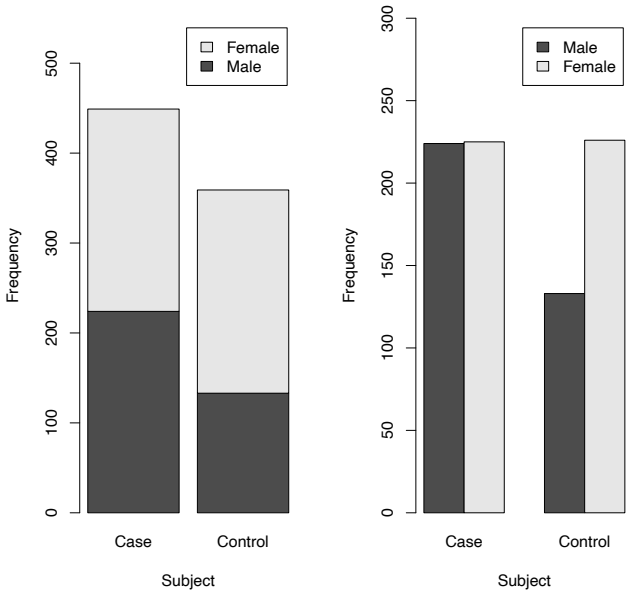

Figure 3: Case and Control Gender Frequency

395 of the cases are current and ex-smokers while 199 of the controls are current and ex-smokers. Based on the Questionnaire regular smokers is defined as smoking as many

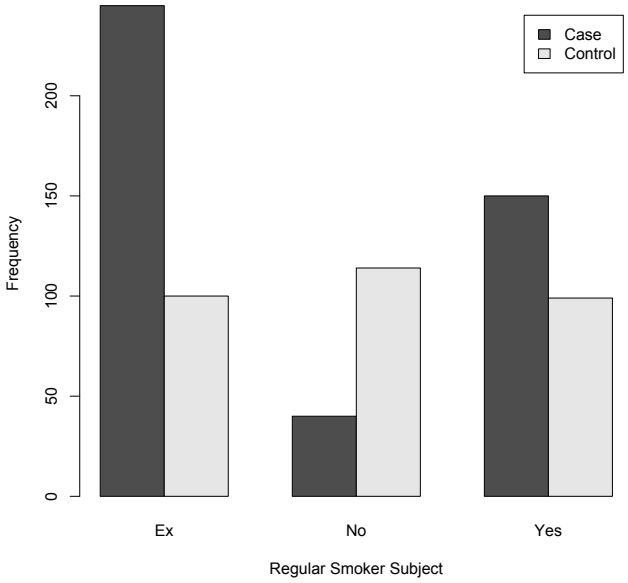

Figure 4: Regular Smoker (10 cigarettes per week for a year)

as 10 cigarettes per week for a year. Figure 4 shows a larger number of cases are regular current and ex-smokers while a smaller number of controls are (were) regular smokers. This results in ascertainment bias as the sample is collected in such a way that some members of the intended population are less likely to be included than others. This is a non-random sample of a population in which all individuals were not equally likely to have been selected. This needs to be accounted for, or results can be erroneously attributed to the phenomenon under study rather than the method of sampling.

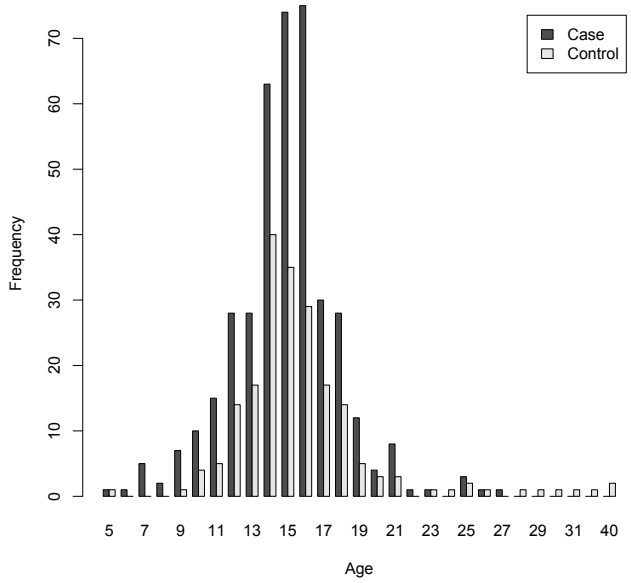

Figure 5: Age At First Cigarette

Plot 5 portrays the starting smoking age with mean 15.23 for both cases and controls. It is evident that for all age groups number of cases starting to smoke is larger than controls.

Plot 6 portrays starting age for regular smoking (overall mean age 17.11). A similar pattern as before is evident among cases with higher starting rate frequency than controls.

Plot 7 reveals that the largest number of ex-smokers attempt to quit at age 50 (S31 mode). Overall mean smoking cessation age is 47.07, patient's mean cessation age is 41.38 years old while case's mean cessation age is 49.4.

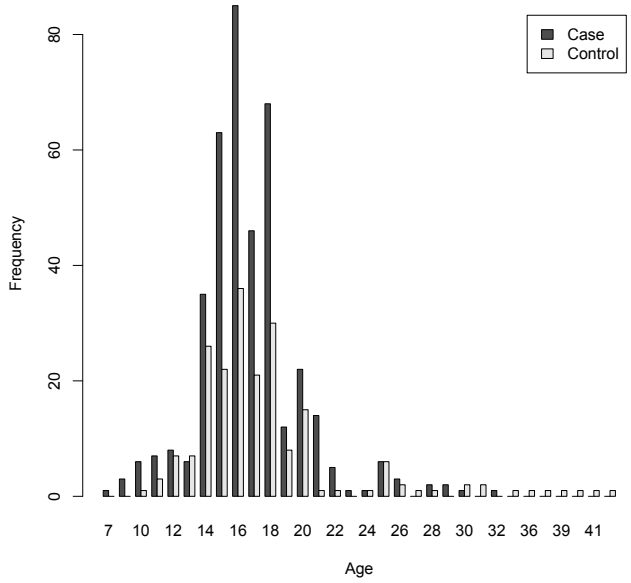

Figure 6: Regular Smoker Starting Age (10 cigarettes per week for a year)

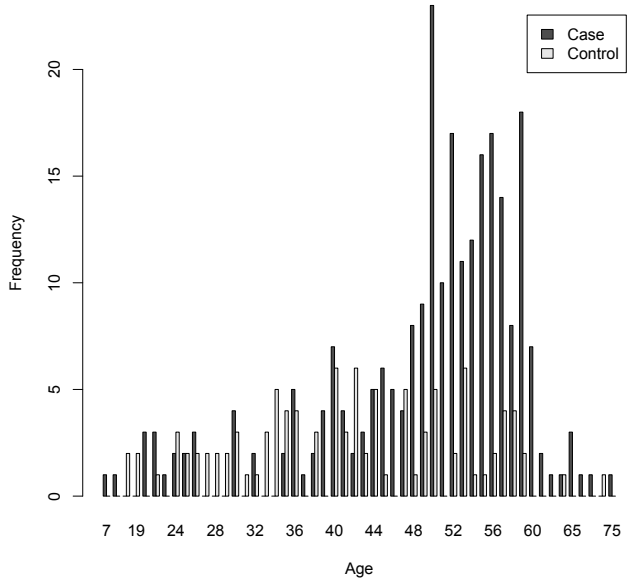

Figure 7: Age Ex-smokers Ceased Smoking

Smoking cessation duration is calculated by subtracting age registered by age of quitting. Excluding NAs in the array (current smokers and non-smokers) results in a

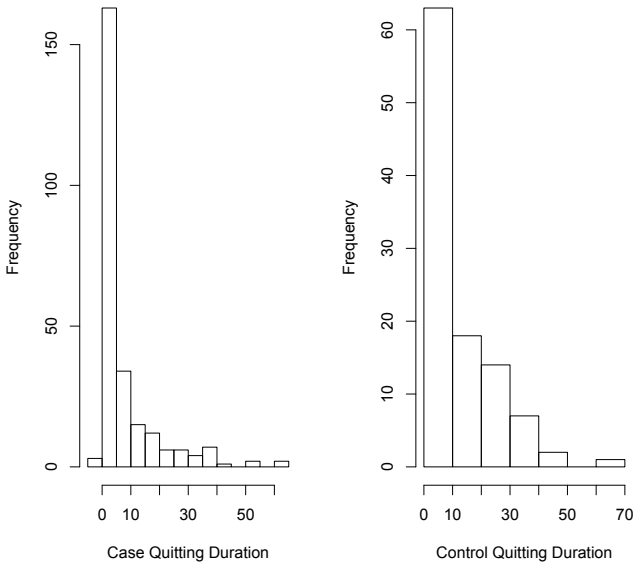

Figure 8: Ex-smokers Smoking Cessation Duration

mean of 29.75 smoking cessation for all ex-smokers. Mean smoking cessation for cases is 7.253 and for controls 11.49. Figure 8 is a histogram of smoking cessation duration. This shows that a larger number of cases and controls quit smoking for about 10 years. This is a larger number for cases than controls.

Mean smoking duration for smokers (current and ex-smokers) is 32.34 years among regular smokers. This is calculated by subtracting age stopped smoking by age started smoking regularly for ex-smokers and current age (at registration) minus age started smoking for current smokers. The results and the plot excludes "Nil" subtraction values ie. non-smokers are excluded from these statistics. Mean smoking duration is 31.84 years among 322 cases and 32.94 years among 275 controls. Plots 9 are bell shaped frequency distributions of smoking duration for cases and controls.

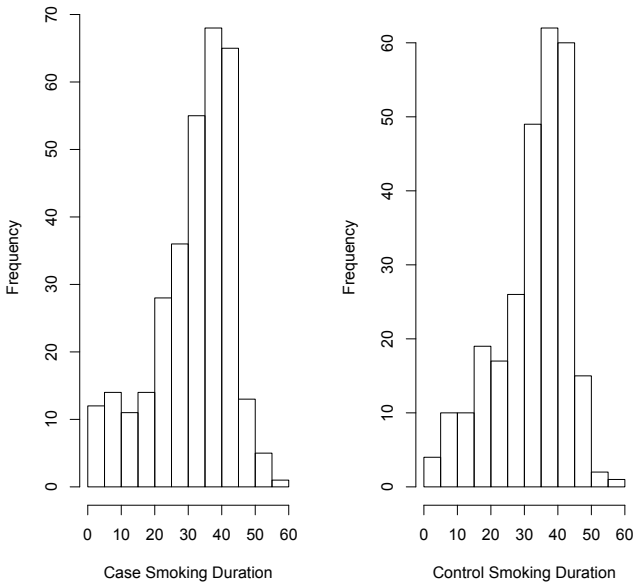

Figure 9: Smoking Duration

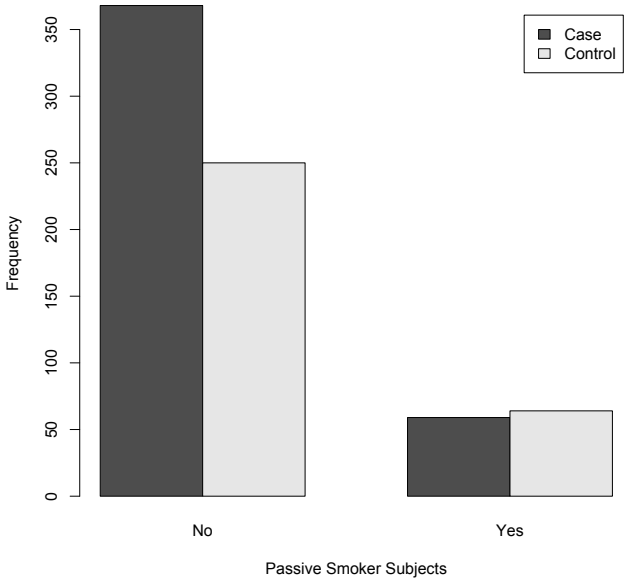

Figure 10: Passive smoking Frequency

Plot 10 shows the frequency of passive smokers among controls and cases. 14% of cases and controls were passive smokers, 70% were not passive smokers while 16% were unknown. There is evidence that passive smoking may not lead to a higher rate of cancer.

Plot 11 shows the frequency of asbestos exposure among controls and cases. As evident from the bar charts there is no evidence that asbestos leads to a higher rate of lung cancer.

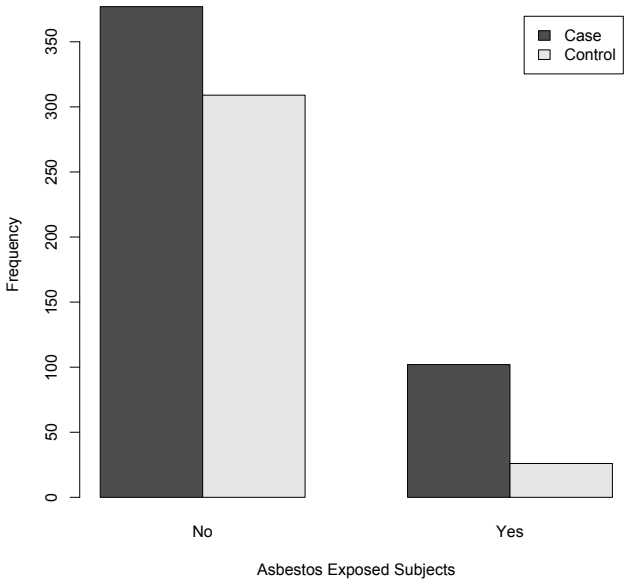

Figure 11: Asbestos Exposure

Miners or factory workers are exposed to various sources of dust (brick, land, steel, fabric, fibreglass, fabric and sawdust). A total of 40% of subjects in the data set were exposed to dust but 52% were not exposed to dust through current and past occupation. Plot 12 is a bar plot of dust exposure for cases and controls. 237 of cases were exposed to dust while only 116 of controls were exposed to dust. There seems to be an association between dust exposure and lung cancer.

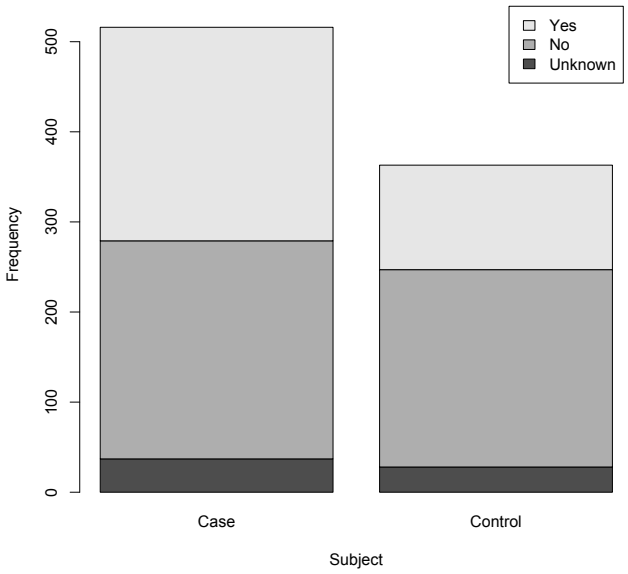

Figure 12: Dust Exposure

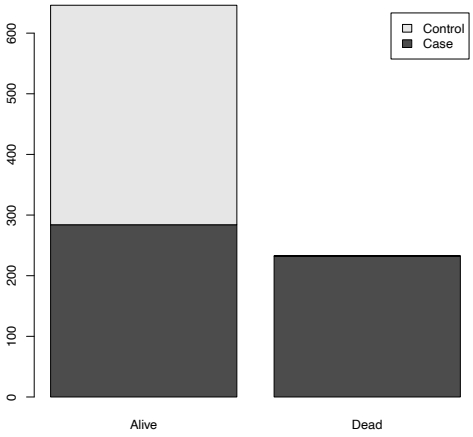

Figure 13: Subjects Status

Plot 13 shows the status of subjects by the end of study. The number of cases alive is 284 (73 %) while the number of controls alive is 362 (27%). Average time alive after diagnosis is 12.85 months.

Plot 14 shows cigarette types at ages 20, 30, 40 and 50. Filtered cigarettes are more common among cases than controls while the rate of cases smoking non-filtered cigarette is higher among cases than controls.

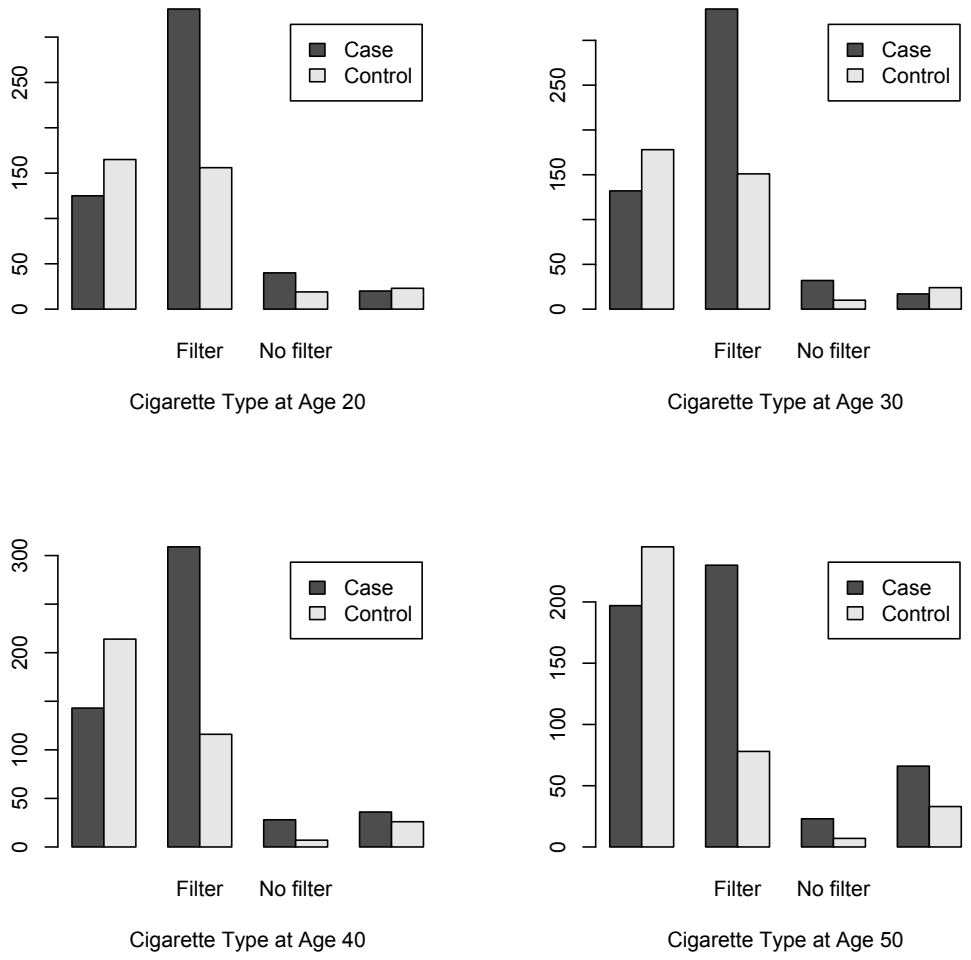

Figure 14: Cigarette Type

Pie chart 15 portrays occupation percentage for cases. This is based on self reported free text box in the questionnaire. Subjects could report as many occupations as they had. 18% are in Routine and Semi-Routine manual and service occupations. Risk models vary for different occupations, for instance miners who are not allowed to smoke during working hours have a different risk model.

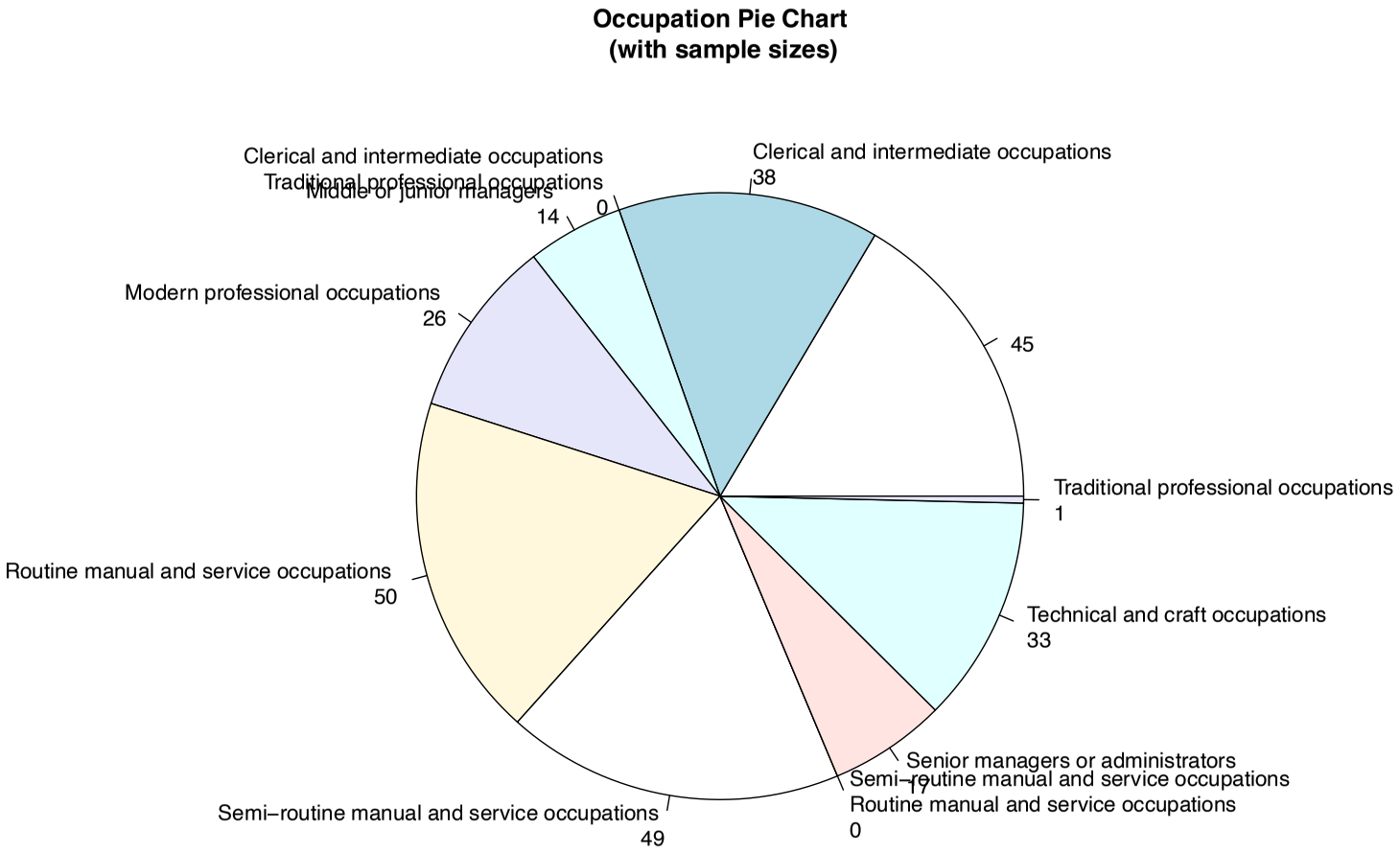

Figure 15: Occupation Pie Chart

Figure 16 shows the lung site affected. In ReSoLuCENT data 28% of right lungs were affected by cancer while 21% of left lungs were. The remaining 51% are controls or unknown.

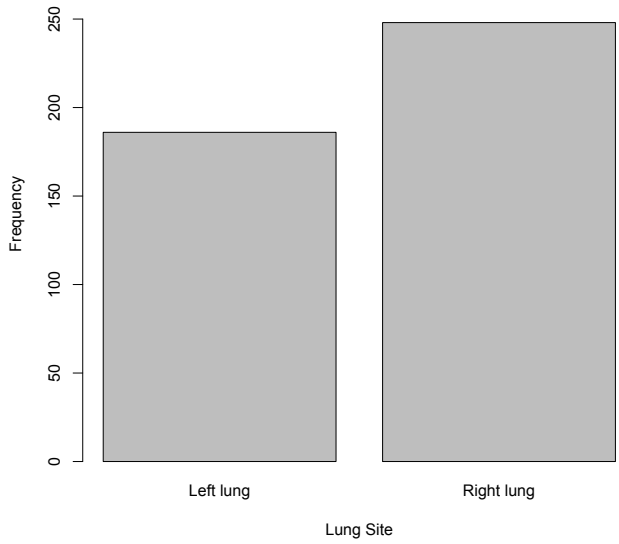

Figure 16: Lung Site

3.3 Family History

In the D'Amelio study, participants were classified as positive for a family history of any smoking-related cancer if at least one first-degree relative had had cancer at some point in his or her life.

In the LLP risk model, information on history of cancer among first degree relatives (i.e. parents, brothers and sisters and biological children) was recorded, including age of diagnosis, site of cancer and relation to the participant. There was a significant increasing trend of risk with numbers of affected relatives, but there was no significant effect of family history (any vs none) of lung cancer in the study population overall or in late-onset cases, regardless of the age of affected relatives. However, there was a substantial and statistically significant increase in risk where both the lung cancer case and the affected relative were diagnosed with lung cancer before the age of 60 years.

The family history of cancer or lung cancer was a statistically significant predictor of lung cancer risk in previous studies. These studies have shown that a first-degree family history of lung cancer is associated with lung cancer risk. This association could be explained by shared genes, shared smoking patterns, or both. Increased risk of lung and other smoking-related cancers among first-degree relatives of lung cancer patients provides evidence for the contribution of both exposure and genetic susceptibility to risk. In this study smoking-related cancer is assigned as renal cancer and cancers of the lung, upper aerodigestive tract, esophagus, pancreas, bladder, and cervix.

Data on cases' family history is provided in Health sheet of ReSoLuCENT data. Data comprises of age, smoking status, underlying health conditions such as lung and heart diseases as well as any other cancer and relevant treatment received. Descriptive statistics is based on data in Table 3. As more than 1 family member can be interviewed, there is a possibility of double counting a case. Although this will not affect risk prediction model.

Number of cases' and controls' family members (parents and siblings) alive were 3872 (60%) ,while 1292 (20%) were dead.

Ex-smokers were 1123 (17%) of the sample population, while 2111 (33%) were non-smokers and 1866 (29%) were current smokers.

5311 (82%) of family history data had no cancer and 901 (14%) had cancer. 85 (1%) had a second cancer. Lung is the major body part affected in this data with 297 subjects, breast 96 and stomach 44 subjects.

Record "H1Cause" in Health data sheet contains information on cause of death of the subject as well as "ID" of cases and controls. For instance for ID number: "100110" medical history, treatments and other health related information of case him/her self, father, mother, brother, sister and child has been reported. To investigate the family history of lung cancer, sufferers were identified manually on Health sheet; the relevant number of sufferers for each case and control was transferred to a column in Main data sheet. Bronchial carcinoma (lung cancers that begins in the tissues of the bronchi, or breathing tubes, inside the lungs) is included in these statistics. 134 of case's relatives were diagnosed with lung cancer. 91% of lung cancer sufferers had 1 case of lung cancer in the family while 9% had 2 cases of lung cancer in their family history. 99 cases were reported to have lung cancer hence 35 controls. 99 (70%) of cases diagnosed with lung cancer had a father who also had lung cancer while 42 (30%) had a mother who had had lung cancer.

Smoking increases the risk of other cancers as well as lung cancer. The Cancer Research UK Web page [4] describes other smoking related cancers as the mouth, larynx (voice box), pharynx (upper throat), nose and sinuses, oesophagus (food pipe), liver, pancreas, stomach, kidney, bladder, cervix and bowel, as well as one type of ovarian cancer and some types of leukaemia. There is also some evidence that smoking could increase the risk of breast cancer. In ReSoLuCENT data approximately 5500 died of heart failure while 145 died of lung cancer, 21 Emphysema, 7 of throat cancer, 4 liver failure or cirrhosis, 12 stomach cancer, 4 kidney failure, 2 bladder cancer, 14 bowel cancer, 4 ovarian cancer and 10 leukaemia and 9 breast cancer. Furthermore 6 died of brain tumour. 5 died of pneumonia, 5 died of Alzheimer's and 2 of prostate cancer.

4 Variable Table

The standardised lifestyle questionnaire collected detailed information on socioeconomic and demographic characteristics, medical history, family history of cancer, history of tobacco consumption and lifetime occupational history. Extensive information about tobacco smoking was elicited for all participants including smoking status, inhalation, type of cigarette smoked, number of cigarettes smoked per day at ages 20, 30, 40 and 50 , age at start and end.

Inclusion of risk factors in the three lung cancer risk models is summarised in Table 4. All these variables were measured in ReSoLuCENT except for hay fever and pneumonia. Wood dust exposure is questioned generally as dust exposure only. Variables below were used for the purpose of calculating risk models.

Regular smokers were defined as those who had smoked as much as 10 cigarettes per week for a year. A former smoker was one who had quit smoking at the time of questionnaire for both patients and controls.

Smoking duration was determined by subtracting the age at which the participant had started smoking from either the age at which the participant had quit smoking (former smokers) or the participant's current age (current smokers).

Pack years were calculated by multiplying the smoking duration (in years) by the number of cigarettes smoked per day and then dividing by 20.

Time of smoking cessation for former smokers was determined by subtracting the age at which the participant had quit smoking from the participant's current age.

Participants were classified as positive for asbestos exposure if they had been exposed or if they were employed in an asbestos-related industry.

Exposure to dusts (including wood dust, sawdust or sanding dust) was self-reported.

Any study participant with missing data for any of the risk factors for any model was dealt with by exclusion or assuming to have a neutral effect on analysis.

Variables	Bach	LLP
Cigarettes smoked per day	Yes	No
Smoking Duration	Yes	Yes
Pack years	No	No
Cessation duration	Yes	No
Age stopped smoking	No	No
Age	Yes	LC incidence rate
Sex	Yes	LC incidence rate
Family history	No	Yes
Asbestos exposure	Yes	Yes
Wood dust exposure	No	No
Emphysema	No	No
Hay fever	No	No
Pneumonia	No	Yes
Malignant tumour	No	Yes
LC incidence rate	Yes (1-year recursed 5 times)	Yes
LC-free mortality rate	Yes (1-year recursed 5 times)	No

Table 4: Lifestyle variables used in the Bach, LLP and ReSoLuCENT Model. Abbreviations: LLP = Liverpool Lung Project; LC = lung cancer; SEER = Surveillance and End Results; NCHS = National Center for Health Statistics. [3]

4.1 Results Table

Table 5 represents epidemiological and demographic data for ReSoLuCENT data.

Cases were older than controls with mean 55.27 compared to mean 51.27. They also had higher mean pack year 26.49 versus 13.62 for controls.

150 cases were currently a smoker while 99 controls were current smokers, although larger number of cases had quit. 245 of cases had quit smoking while only 100 of controls had quit.

For current smokers smoking duration was higher in cases than controls while this was reversed in former smokers. Pack year was only slightly different for current and ex-smokers. There is a slight increasing trend with age in number of cigarettes smoked per day for both current and former smokers. This is evident in both cases and controls.

A larger number of cases had family history of lung cancer compared to controls ie. they had at least one first-degree relative had had cancer at some point in his or her life. 19% had 1 or 2 cases of lung cancer in their family history while 10% of cases had 1-2 family members with lung cancer.

Variables	*Cases*($N = 516$)	*Controls*($N = 363$)
Age (years) : mean ± s.d.	55.27 ± 6.914	51.27 ± 11.03
Pack years : mean ± s.d.	26.49 ± 21.59	13.62 ± 18.02
Smokoing status (no. %)		
Current	150 (0.29)	99 (0.27)
Former	245 (0.47)	100 (0.27)
Current smokers		
Smoking duration (years) : mean ± s.d.	40.37 ± 12.98	38.88 ± 11.36
Cigarettes per day (Age : 20) : mean ± s.d.	17.1 ± 8.66	15.64 ± 7.53
Cigarettes per day (Age : 30) : mean ± s.d.	19.86 ± 10.02	19.69 ± 8.89
Cigarettes per day (Age : 40) : mean ± s.d.	21.68 ± 9.45	20 ± 8.36
Cigarettes per day (Age : 50) : mean ± s.d.	22.78 ± 11.19	19.89 ± 9.39
Pack years: mean ± s.d.	26.83 ± 19.94	13.57 ± 19.07
Former smokers		
Smoking duration (years) : mean ± s.d.	31.84 ± 11.6	32.94 ± 10.73
Cigarettes per day (Age : 20) : mean ± s.d.	16.01 ± 7.778	15.6 ± 8.97
Cigarettes per day (Age : 30) : mean ± s.d.	20.57 ± 9.53	20.56 ± 7.77
Cigarettes per day (Age : 40) : mean ± s.d.	22.84 ± 11.2	20.72 ± 9.44
Cigarettes per day (Age : 50) : mean ± s.d.	21.63 ± 10.43	22.69 ± 11.99
Pack years: mean ± s.d.	26.72 ± 22.6	13.29 ± 17.56
Family history of lung cancer (no.% of family members)		
No	417 (80 %)	328 (90%)
1	92 (18 %)	30 (8.2%)
≥2	7 (1.3%)	5 (1.3%)
Early onset ≥ 60 years old	15 (3%)	9 (2%)
Late onset ≤60 years old	84 (2%)	26 (7%)

Table 5: ReSoLucent Series Summary Table

5 Pack Years

For the purpose of statistical analysis and risk model building, pack year variable is used as the measure of smoking intensity for current and former smokers.

Pack year is a measure of cigarette smoking over someone's lifetime, figured as the number of packs per day times the number of years a person has smoked. Ten pack years could refer to a smoking history of two packs a day for five years, one pack a day for 10 years, or half a pack a day for 20 years. 1 pack has 20 cigarettes. One "pack year" means 7300 cigarettes, or 1460 cigars, or 7.3kg of pipe tobacco.

In this project pack year is calculated using the simple formula below:

Number of pack years = (number of cigarettes smoked per day x number of years smoked)/20

In ReSoLuCENT data variables "S31_20", "S31_30", "S31_40" and "S31_50" represent number of cigarettes smoked per day at ages 20, 30, 40 and 50 respectively. If the person has not smoked at any age, there is no value for that specific cell. Code below in R assigns value "zero" to all the missing values for these vectors to simplify the pack year calculation. Furthermore "S30" (age started smoking regularly), "S31" (age stopped smoking) and "AgeRegistered" (age at registration) are assigned a value of 0 for any missing value.

```
#Assign "Nil" to unknown values

for(i in 1:length(S31))

{if (is.na(S30[i])) {S30[i]=0} #Non-Smokers

{if (is.na(S31[i])) {S31[i]=0}            #Current Smokers

{if (is.na(AgeRegistered[i])) {AgeRegistered[i]=0}   #Age joined the study

{if (is.na(S31_20[i])) {S31_20[i]=0}    #Non smoker at age 20
```

```
{if (is.na(S31_30[i])) {S31_30[i]=0}    #Non smoker at age 30

{if (is.na(S31_40[i])) {S31_40[i]=0}    #Non smoker at age 40

{if (is.na(S31_50[i])) {S31_50[i]=0}}}}}}}    #Non smoker at age 50
```

Using the pack year formula above, multiplying each 10 year smoking band by number of cigarettes smoked per day and dividing the total by 20 results in the smoking pack year for each individual. As years that current or ex-smokers did not smoke is now 0 multiplying it by 10 does not add up to pack years. "S31_20" is assumed to be number of cigarettes smoked in age band 20-30, "S31_30" age band 30-40, "S31_40" age band 40-50 and "S31_50" age band 50-60.

```
( 10 * S31_20[i] + 10 * S31_30[i] + 10 * S31_40[i] +  10 * S31_50[i] ) /20
```

A more complex algorithm is used to calculate more accurate pack year excluding years that a subject did not smoke in between each 10 year band or to include years that subjects smoked before age 20 or after age 60. For instance if a smoker stopped smoking after age 50, smoking cessation age (S31) is subtracted by 50 then multiplied by number of cigarettes smoked per day at age 50 (S31_50). If a subject started smoking before 20, then 20 is subtracted by starting smoking age (S30) and multiplied by the number of cigarettes smoked at age 20 (S31_20).

Non-smokers never start smoking regularly (S30=0) and obviously never stop smoking either (S31=0).

Current smokers have not stopped smoking ie. (S31=0). Calculating pack years depends on subject's current age as they are still smoking. Using several if statements this age has been located in a specific age band, more than 50, 40, etc. and multiplied by relevant number of cigarettes smoked per day.

An ex-smoker's pack year is calculated depending on age they stopped smoking whether over 50, 40 etc.

The first for loop calculates pack years excluding starting smoking ages in bands 20-30, 30-40 and 40-50. For instance if the subject started smoking at age 25, as the number of cigarettes smoked per day at age 20 (S31_20) is 0; the 5 years smoking (25-30) will not be included in the pack year. To include this, the number of cigarettes per day for age 30 is extrapolated to age 20-30 band and added to the calculated pack year in the second for loop. Second code should only run once to avoid adding up to the already calculated pack years.

Diagram 17 illustrates age with respect to number of cigarettes smoked per day for subject 141 of ReSoLuCENT data. This is an ex-smoker who started smoking at age 18, and stopped smoking at age 55. As evident from this diagram number of cigarettes smoked per day before age 20 is assumed to be 20; same as that for age band 20-30 (S31_20).

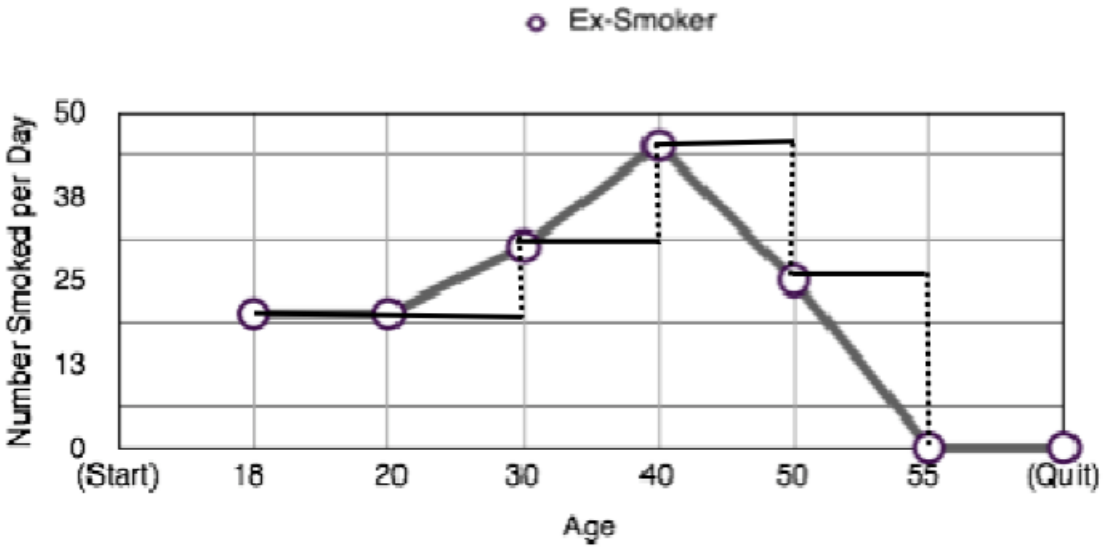

Figure 17: Pack Years diagram for subject i=141

Variables	Ex-Smoker
i-th Subject	141
S31[i] Cessation Age	55
S30[i] Starting smoking regularly	18
S31_20[i] Cigarettes per day (Age : 20)	20
S31_30[i] Cigarettes per day (Age : 30)	30
S31_40[i] Cigarettes per day (Age : 40)	45
S31_50[i] Cigarettes per day (Age : 50)	25

Table 6: Pack Year Calculation for subject i=141 in Figure 16

```
p<-rep(NA,length(S31)) #Create an array of NAs with length S31

for(i in 1:length(S31))
{if (S30[i]==0 && S31[i]==0) {p[i]=0}   # Non-smokers

else
{if ((S31[i]==0) && AgeRegistered[i]>=50)
# Current smoker ,current age over 50
{p[i]<-((AgeRegistered[i]-50)*S31_50[i]
+10*S31_40[i]+10*S31_30[i]+10*S31_20[i]
+((20-S30[i])*S31_20[i]))/20}

else
{if ((S31[i]==0) && AgeRegistered[i]>=40)
# Current smoker ,current age over 40
{p[i]<-((AgeRegistered[i]-40)*S31_40[i]
+10*S31_30[i]+10*S31_20[i]
+((20-S30[i])*S31_20[i]))/20}
```

```
else
{if ((S31[i]==0) && AgeRegistered[i]>=30)
# Current smoker ,current age over 30
{p[i]<-((AgeRegistered[i]-30)*S31_30[i]
+10*S31_20[i]+((20-S30[i])*S31_20[i]))/20}

else
{if ((S31[i]==0) && AgeRegistered[i]>=20)
# Current smoker ,current age over 20
{p[i]<-((AgeRegistered[i]-20)*S31_20[i]
+((20-S30[i])*S31_20[i]))/20}

else
{if ((S31[i]==0) && AgeRegistered[i]>=15)
# Current smoker ,current age over 15
{p[i]<-((AgeRegistered[i]-S30[i])*S31_20[i])/20}

else
{if (S31[i]>=50)   # Stopped smoking after 50
{p[i]<-(((S31[i]-50)*S31_50[i])+10*S31_40[i]+10*S31_30[i]+10*S31_20[i]
+((20-S30[i])*S31_20[i]))/20}

else
{if (S31[i]>=40)  #  Stopped smoking after 40
{p[i]<-(((S31[i]-40)*S31_40[i])+10*S31_30[i]+10*S31_20[i]
+((20-S30[i])*S31_20[i]))/20}

else
{if (S31[i]>=30)  #  Stopped smoking after 30
{p[i]<-(((S31[i]-30)*S31_30[i])+10*S31_20[i]
+((20-S30[i])*S31_20[i]))/20}

else
{if (S31[i]>=20)  # Stopped smoking after 20
{p[i]<-((S31[i]-20)*S31_20[i]+((20-S30[i])*S31_20[i]))/20}
```

```
else
{if (S31[i]>=17)
{p[i]<-((S31[i]-S30[i])*S31_20[i])/20}  # Stopped smoking after 17

}}}}}}}}}}

# 2nd For Loop: Includes starting smoking age in bands
20-30, 30-40 and 40-50

for(i in 1:length(S31))  # Started smoking in age bands 20-30
{if (p[i]>0 && 30 >S30[i] && S30[i]>20)
{p[i]<-p[i]+((30-S30[i])*S31_30[i])/20}

else # Started smoking in age bands 30-40
{if (p[i]>0 && 40 >S30[i] && S30[i]>30)
{p[i]<-p[i]+((40-S30[i])*S31_40[i])/20}

else # Started smoking in age bands 40-50
{if (p[i]>0 && 50 >S30[i] && S30[i]>40)
{p[i]<-p[i]+((50-S30[i])*S31_50[i])/20}
}}}
```

This code can be validated by hand calculation for an ex-smoker, current smoker and non-smoker in Table 7. Pack year for subject i (p[i]) is calculated by R algorithm for each subject and is validated using manual calculation.

Figure 18 is a boxplot of smoking pack years for all ReSoLuCENT subjects, controls and cases. As evident from the plot the median smoking pack year for cases is 26.95 and considerably higher than median for controls at 4.14. This is 19.25 for all subjects. A large outlier is observed among cases, this is for subject i= 823 at pack year = 121.1.

There are 2 NA values observed in the pack year calculation. For i=835 where starting smoking age (S30) is larger than age stopped smoking age (S31) and i=431 where there is no date of birth. Boxplot and basic pack year statistics are calculated including these NA values but excluding NA values from subset "case pack years" and "control pack years".

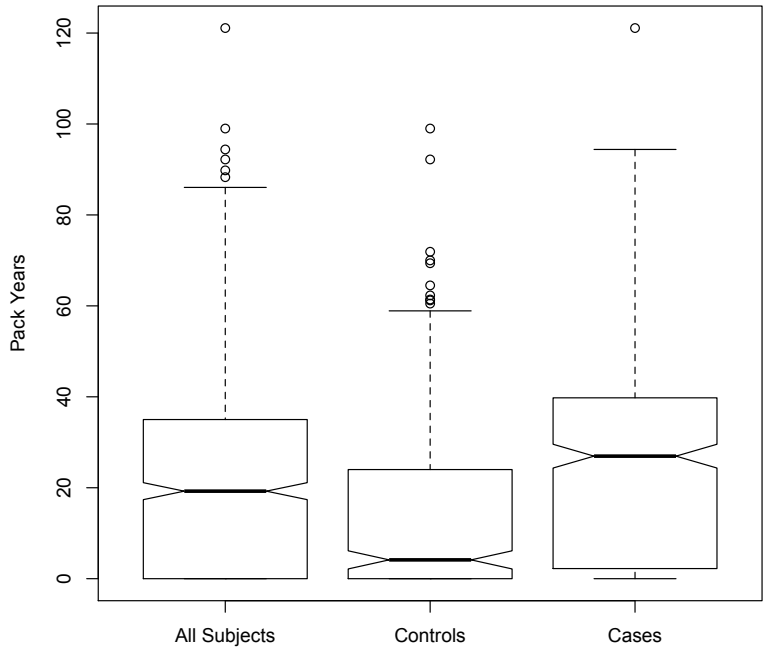

Figure 18: Pack Years=(Number of cigarettes smoked per day x years smoked) /20

Variables	Non-Smoker	Ex-Smoker	Current Smoker
i-th Subject	2	50	4
S31[i] Cessation Age	0	56	0
S30[i] Starting smoking	0	10	15
S31_20[i] Cigarettes per day (Age : 20)	0	20	15
S31_30[i] Cigarettes per day (Age : 30)	0	20	15
S31_40[i] Cigarettes per day (Age : 40)	0	25	0
S31_50[i] Cigarettes per day (Age : 50)	0	25	0
p[i] Pack years	0	50	18.07
Hand Calculation	0	((20-10)*20 + 20*10 + 20*10 + 25*10 + 25*(56-50)) /20	Age: (30-39.1) + (20-30) + (15-20) ((39.1-30)*15 + 10*15 + 5*15)/20

Table 7: Pack Year Validation Table

6 Risk Prediction

Risk Assessment Models estimate the binary outcome of developing cancer over a defined period of time. Cancer risk prediction models help identify high risk individuals and control suffering and death by preventative methods. Models also facilitate the design and planning of clinical chemoprevention trials, development of benefit-risk indices and provide estimates of the population burden and cost of cancer. Lung cancer model inputs are : subject's gender, sex, family history, asbestos exposure history and smoking history (cigarettes per day, smoking duration, smoking cessation duration).

6.1 Comparison Methods

Using variables mentioned, absolute risk for lung cancer (one from each model), for each participant is estimated. Specificity and sensitivity is calculated for each model to construct ROC curves and estimate AUC (binomial method). Models' discriminatory power has been compared by calculating the area under the curve (AUC) of the receiver operator characteristic (ROC) curve. The discriminatory power, accuracy, positive predictive value (PPV; the probability of accurately categorising an affected participant) and negative predictive value (NPV; the probability of accurately categorising an unaffected participant) and clinical utility of these models have been compared.

Sensitivity is the proportion of times the test is positive given the patient has the disease, ie. 1 - proportion of false negatives (test is negative when they have disease). When a result is negative, tests with high sensitivity are useful for ruling out a diagnosis.

Given a subject is free of disease, specificity is the proportion of times a test will be negative , ie. 1 - proportion of false positives (i.e. test is positive when person does not have disease). When a test is positive, tests with a high specificity rule in a diagnosis.

Both sensitivity and specificity are conditional on having or not having the disease. They are properties of the test, not the disease. While positive (and negative) predictive values depend on prevalence.

Plot ROC (Receiver Operator Characteristics) Curve is sensitivity versus 1 minus specificity for various cut-off points. For continuous test of sensitivity versus 1 minus specificity a cut-off is decided above which the test is positive. The best cut-point is the

one nearest the upper left-hand corner. When comparing two curves the one with the largest area under the ROC is considered more accurate.

The determination of an "ideal" cut-off value is a trade-off between sensitivity (true positives) and specificity (true negatives). The ROC curve offers a graphical illustration of these trade-offs at each "cut-off" for any diagnostic test that uses a continuous variable. Ideally, the best "cut-off" value provides both the highest sensitivity and the highest specificity, located on the ROC curve by finding the highest point on the vertical axis and the furthest to the left on the horizontal axis.

The area under the ROC curve (AUC) is widely recognized as the measure of a diagnotic test's discriminatory power. The maximum value for the AUC is 1.0, indicating a (theoretically) perfect test (i.e. 100% sensitive and 100% specific). An AUC value of 0.5 indicates no discriminative value (i.e., 50 % sensitive and 50% specific) and is represented by a straight, diagonal line extending from the lower left corner to the upper right. A ROC curve that is no better than chance will lie along 45 °line. ROC analysis provides important information about diagnostic test performance: the closer the apex of the curve toward the upper left corner, the greater the discriminatory ability of the test (i.e., the true-positive rate is high and the false-positive [1 - Specificity] rate is low).

The 95% AUC confidence interval is computed with Delong's method (DeLong et al.) based on U-statistics theory and asymptotic normality as this test does not require bootstrapping. [6]

6.2 Bach Model

The Bach lung cancer risk prediction model was derived using a database of 18,172 subjects enrolled in the Carotene and Retinol Efficacy Trial (CARET) a large, randomized trial of lung cancer prevention [1]. The extent of variation in risk was assessed in a cohort of individuals who met typical eligibility criteria for cancer prevention studies.

The Bach model estimates the absolute risk that an individual will be diagnosed with lung cancer within 10 years. To determine the absolute risk of lung cancer for an individual within 10 years, two 1-year multivariate models were created. One predicts the probability of being diagnosed with lung cancer (the focus of project), and the other predicts the probability that an individual will die without having been diagnosed with

lung cancer (the competing risk). 10 year lung cancer risk was then recursively estimated by cycling these two 1-year models 10 times. In each year, the risk of lung cancer diagnosis and the risk of death in the absence of lung cancer were estimated. For each year two scenarios were taken into account: continued smoking (at the same level) for current smokers and continued abstinence from smoking for ex-smokers.

Study entry criteria were mainly for at-risk subjects, aged 55 - 74 years, who had smoked a minimum of 30 pack years, and current smokers or former smokers who quit within the last 15 years. Models calculate the 1-year probability of diagnosis of lung cancer and the 1-year probability of survival in the absence of lung cancer. Model inputs such as subject's age, duration of abstinence, duration of smoking, and number of cigarettes smoked per day were treated as continuous predictors while sex, and asbestos exposure were treated as categorical variables. The individual's age at the time he or she started smoking is included in predictors such as duration of smoking, and duration of abstinence, and therefore is not directly used in the analysis. CPD (cigarettes per day) variable was calculated from pack year variable as this is average number of cigarettes smoked per day throughout individual's smoking duration. For instance a person who started smoking at 15, smoked 20 cigarettes per day at the age 20 and 30, 35 cigarettes per day at age 40 and stopped smoking when 44 years old has CPD of 22.07.

Using cumulative distribution function of Weibull distribution; one-year probability of death in the absence of a diagnosis of lung cancer is : 1 - S_0 * (e^{-model}) ,where S_0 = 0.9917663 and "model" is estimated based on Bach lung cancer prediction equations in figure 21 in the Appendix. For one-year probability of a diagnosis of lung cancer S_0 = 0.99629. For the purpose of comparing results with that of D'Amelio study 5 year risks have been calculated for ReSoLuCENT data. Cycling 1 year diagnosis and death probability models 5 times, 5 year lung cancer risk is estimated recursively. 5 year risk is accumulative risk of diagnosis on that year, on the condition of survival and no diagnosis of lung cancer and in previous years. For instance:

Lung cancer risk prediction in Year 2 =
Pr. (Diagnosis in Year 2) * (1-Pr. (Diagnosis in Year 1))* (1 - Pr. (Death in Year 1))

The 5 year additive risk evaluated by this recursive model is then validated against the online risk assessment tool shown in figure 23 from the Journal of National Cancer

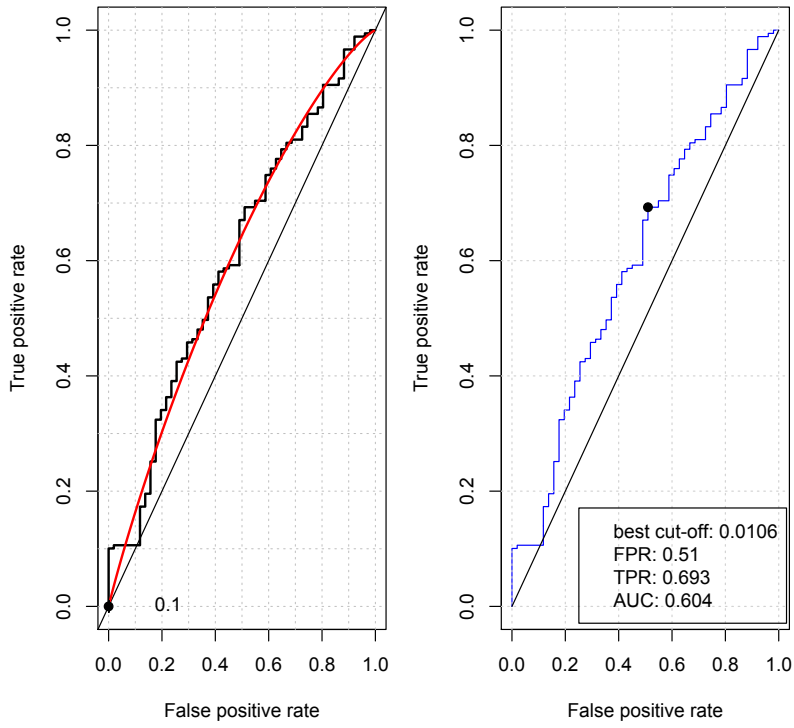

Figure 19: Receiving Operating Curve for the Bach risk model

Institute (JNCI) [9]. For instance the 10 year risk of developing lung cancer for ReSoLuCENT's subject 34 who is a 64 year old male who smoked 44 cigarettes per day for 17 years and was not exposed to asbestos is 7% assuming he continues to smoke at the same rate. Using recursive models in R software the 5-year risk of cancer was calculated as 1.05%. This seems slightly lower than anticipated, I require further time to check details of rigorous analysis in case of any mistake, another explanation may be that ReSoLuCENT data has a different format to CARET data. The mean 5 year risk of lung cancer is 1.37%. The maximum risk is 7.37% for subject 508 a 56.8 year old female case with 20 CPD for smoking duration of 28 years.

Figure 19 is ROC curve of Bach model on ReSoLuCENT data with AUC of 0.60. For the best cut-off point of 1.04%, sensitivity is 0.69 and specificity is 0.49. For cut-off 2.5% (as in D'Amelio) , sensitivity is 0.88, specificity is 0.16, PPV is 0.23 and NPV is 0.82.

6.3 Liverpool Lung Project Model

In this study absolute risk of lung cancer is calculated over a defined period, based on data from a case control study, the Liverpool Lung Project (LLP) (Field et al, 2005) [2]. Participants were 579 cases of lung cancer and 1157 population controls recruited between 1998 and 2005 among residents of Liverpool area.

Continuous model inputs were subject's age and smoking duration while occupatinal exposure to asbestos and family history of lung cancer were categorical. Estimates of 5-year lung cancer absolute risk were developed using methods similar to those predicting individualized breast cancer risk, Gail et al (1989) and Chen et al (2006), combining relative risk models with local rates for lung cancer incidence. Performing χ^2 test, student's t-test, Wilcoxon rank sum test, conditional logistic regression, backward stepwise regression, pairwise interaction test, the final multivariate model was built up and the logistic regression model below was arrived at :

$$ln\ (\frac{p}{1-p}) = \alpha\ +\ \sum_i\ \beta_i\ x_i \tag{1}$$

where $i = 1, ..., n$ (n=879 in ReSoLuCENT data)
p : Probability of disease
α : log odds of a positive outcome, for control groups
β_i : Log odd ratios (Model coefficients)
x_i : Risk factor for subject i

Rearranging logarithmic formula (1) , p can be expressed as below:

$$p = \frac{1}{1 + exp\ (\ -(\alpha\ +\ \sum_i\ \beta_i\ x_i))} \tag{2}$$

Since conditional logistic regression does not estimate α, this must be estimated separately. Therefore multivariate logistic model is converted to estimate absolute risks. Table 8 shows the derived αs and their standard errors, by age group and sex (taken from Table 24 in appendix). α is then calculated for each individual taking into account the exact age (below or above the age band). β variable in Table 9 (taken from Table 25 of Bach paper) does not contain pneumonia variable, as this information has not been elicited in ReSoLuCENT questionnaire. p is then calculated for ever smoker subjects by plugging in for α and β values in formula (2). For comparison of discriminatory power between the LLP and Bach models, only ever smokers were used as the Bach

Age group	Male α -value	Female α -value
40-44	-9.06	-9.90
45-49	-8.16	-8.06
50-54	-7.31	-7.46
55-59	-6.63	-6.50
60-64	-5.97	-6.22
65-69	-5.56	-5.99
70-74	-5.31	-5.49
75-79	-4.83	-5.23
80-84	-4.68	-5.42

Table 8: Age and sex specific lung cancer estimated α values for 5-year absolute risk

model was developed only for ever smokers (excluded non-smokers).

The mean absolute risk for 93 eligible subjects is 2.02%. The maximum risk is 10.60% for subject 56 who is a 58 year old male with 42 years smoking duration and 2 cases of lung cancer in family history.

Figure 20 shows the receiver operating characteristic curve derived when the model was applied to the ReSoLuCENT case control population. The area under the curve is 0.72 similar to D'amilo's result (0.69) indicating good discrimination between cases and controls. Predefined cutoff of 2.5 % used in D'Amelio would capture 33% of lung cancer cases while including 19% of the controls, giving a sensitivity of 67% and specificity of 81%.

In terms of model accuracy, the low (21%) PPV for LLP model indicates that it can not identify high-risk individuals; the relatively high NPV (93%) indicates that many low-risk individuals will be identified.

Risk Factor/ Category	Model coefficient (β -value)
Smoking duration	
Never	0.00
1-20 years	0.769
21-40 years	1.452
41-60 years	2.507
>60 years	2.724
Occupational exposure to asbestos	
No	0.000
Yes	0.634
Family history of lung cancer	
No	0.000
Early-onset (<60 years)	0.703
Late-onset (>60 years)	0.168

Table 9: LLP β model coefficients

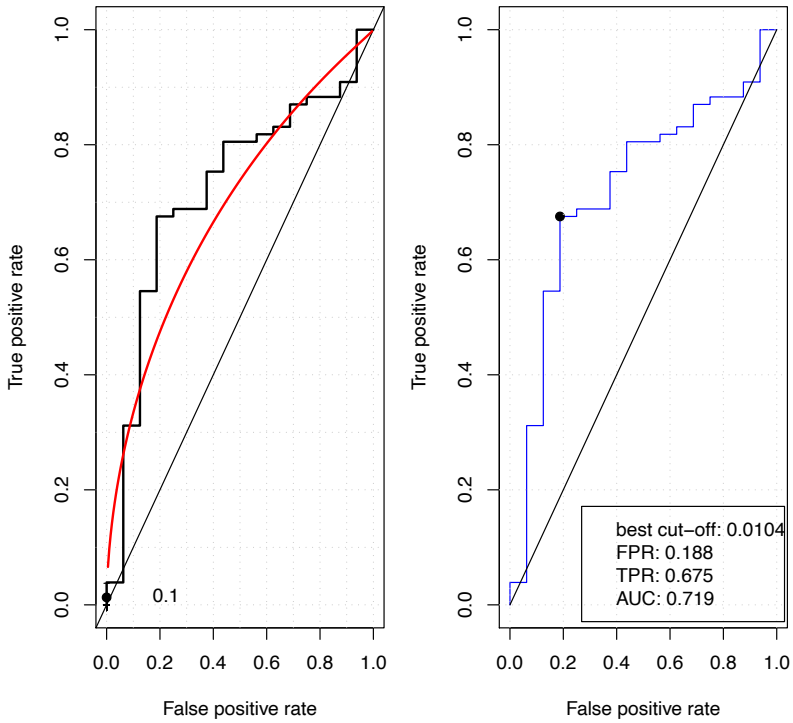

Figure 20: Receiving Operating Curve for the LLP risk model

7 Discussion

The Bach and LLP models have previously been compared in terms of discriminatory power, accuracy, and clinical utility in D'Amelio study on dataset of 3197 patients with lung cancer and 1703 cancer free controls. Each model was used to estimate 5-year absolute lung cancer risks for an independent population of lung caner patients and healthy controls. The discriminatory power of Bach and LLP models for both ReSoLuCENT and D'Amelio study are summarised in Table 10.

Area Under Curve of LLP (0.72, CI: 0.58 - 0.85) is larger than Bach's (0.60, CI: 0.52 - 0.69), therefore it has a better discriminatory power. LLP also has a better discriminatory power in the D'Amelio study. Although the Bach model gives a lower AUC for ReSoLuCENT model (0.60) than D'Amelio (0.66), there is overlap of confidence interval with ReSoLuCENT's 95% CI of (0.64-0.69).

In terms of clinical utility, the LLP model performed reasonably well in identifying

ReSoLuCENT lung cancer patients at defined levels of risk while limiting the number of false-positive results (sensitivity=67% , specificity=81%). However, the Bach model was much better at identifying individuals with lung cancer and had a higher false - positive rate than the LLP model (sensitivity = 88%, specificity=16%). Overall, sensitivity and specificity of the LLP risk model compare favourably with the Bach absolute risk model. In the D'Amelio study, The LLP model had a higher false-positive rate. At all levels of risk, the LLP model correctly identified a higher proportion of lung cancer patients than did the Bach model but also incorrectly identified a higher proportion of controls as lung cancer patients. This could be attributed to the importance of smoking in the LLP model.

In terms of model accuracy, the models do not have reasonable PPV levels (all <70%); Bach had a higher PPV (23%) than the LLP (21%). The overall NPV for each of the models were higher than the PPV, with the LLP model having a substantially better probability of accurately identifying many low risk individuals. D'Amelio's higher PPV result for Bach and LLP model show they have a substantially better probability of accurately categorising an affected participant.

Overall the LLP risk models higher discriminatory power and simplicity makes it more directly applicable for use in the primary care setting.

Model	D'Amelio		ReSoLuCENT	
	Bach	LLP	Bach	LLP
Area under the curve	0.66	0.69	0.60	0.72
Asymptotic 95 % confidence interval	0.64 - 0.69	0.67 - 0.71	0.52 - 0.69	0.58 - 0.85
Sensitivity	<0.62	0.62	0.88	0.67
Specificity	>0.70	0.70	0.16	0.81
PPV	0.81	0.76	0.23	0.21
NPV	0.45	0.56	0.82	0.93

Table 10: Comparison of discriminatory power and accuracy of lung cancer risk models using a predictive cut off of 2.5%

8 Limitations

In ReSoLuCENT study controls are selected from partners or siblings of cases; they may share environmental factors and similar family history.

Risk prediction models have some limitations. They do not distinguish between the risks of different histologic types of lung cancer, and are relevant only to one subset (albeit a large subset) of at risk individuals : those aged 50 years or older who have a smoking history.

In the Bach and LLP model subjects were participants in a clinical trial of lung cancer prevention and therefore not perfectly representative of members of the population at large. A limitation of the LLP risk model is that the absolute risks estimated for each combination of risk factors are based on relative risks derived from a case control study.

Also there is a potential that recall and other information biases could influence the final results, as cases and controls were asked to report their lifestyle habits and behaviours for many years prior to interview. Cancer studies use standardised life style questionnaire to elicit relevant information but have different definitions for ex-smokers, regular smoker, pack year, family history and asbestos exposure.

Another limitation is that the risk prediction models compared in this study were developed in Caucasian populations, so the models may not be applicable to other racial or ethnic groups.

9 Further Study

With more time I would stratify ReSoLuCENT participants by age and sex and conduct pairwise comparisons of the AUCs of models using the method described in the NCSS package (Hanley and McNeil, 1983) [11] to test the differences in discriminatory power between the LLP and Bach model. D'Amelio study compares discriminatory power of the Spitz [10] lung cancer risk model along with the LLP and Bach model. With more time I assess the discriminatory power of the Spitz model on ReSoLuCENT data.

Although the results suggest that the LLP risk model is more useful for predicting risk, further research is needed to test the applicability of the model in diverse populations, including those from diverse geographic regions. A subset of the ReSoLuCENT data has additional laboratory based genetic measures potentially associated with lung cancer risk. The risk prediction models can further investigate if the inclusion of the additional laboratory based genetic measures improves the risk prediction.

10 References

Bibliography

[1] Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, Hsieh LJ, Begg CB (2003) Variations in lung cancer risk among smokers. J Natl Cancer Inst 95: 470478

[2] Cassidy A, Myles JP, van Tongeren M, Page RD, Liloglou T, Duffy SW, Field JK (2008) The LLP risk model: an individual risk prediction model for lung cancer. Br J Cancer 98: 270276

[3] AM DAmelio Jr1, A Cassidy1,2, K Asomaning3, OY Raji2, SW Duffy4 (2010) Comparison of discriminatory power and accuracy of three lung cancer risk models. British Journal of Cancer (2010) 103, 423 429

[4] Cancer Research UK;

http://info.cancerresearchuk.org/cancerstats/mortality/cancerdeaths/

[5] National Cancer Institute : http://riskfactor.cancer.gov/

[6] Elisabeth R. DeLong, David M. DeLong and Daniel L. Clarke-Pearson (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach

[7] Peto, R., et al., Mortality from smoking in developed countries 1950-2000: Indirect estimates from National Vital Statistics. 1994, Oxford: OUP.

[8] Resource for the Study of Lung Cancer Epidemiology in North Trent; http://resolucent.group.shef.ac.uk/index.html

[9] Variations in Lung Cancer Risk Among Smokers; http://jnci.oxfordjournals.org/content/95/6/470/suppl/DC1

[10] Spitz MR, Hong WK, Amos CI, Wu X, Schabath MB, Qiong D, Shete S, Etzel CJ (2007) A risk model for prediction of lung cancer. J Natl Cancer Inst 99: 715726

[11] Hanley JA, McNeil BJ (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148: 839843

[12] Michael J. Crawley (2007) The R book

[13] John Verzani (2004) Using R for Introductory Statistics

11 Appendix

11.1 ReSoLuCENT Team

Chief Investigator Prof Penella J Woll, Department of Clinical Oncology, University of Sheffield

Investigators
Dr M Dawn Teare, Lecturer in Genetic Epidemiology, University of Sheffield
Dr Angela Cox, Senior Lecturer, Institute for Cancer Studies, University of Sheffield

Trials Co-ordinator:
Lesley Turner
(l.a.turner@sheffield.ac.uk)
Tel. 0114 2265219

Clinical Trials Practitioner:
Sue Ellis
(s.p.ellis@sheffield.ac.uk)
Tel. 0114 2265218

Address:
Cancer Clinical Trials Centre
Weston Park Hospital
Whitham Road
Sheffield
S10 2SJ

The one-year probability of death in the absence of a diagnosis of lung cancer = $1-S_0^{e^{(model)}}$

Where,

S_0 = 0.9917663;

CPD = cigarettes per day, SMK = duration of smoking, QUIT = duration of quitting, AGE = age, ASB = asbestos exposure, and SEX = sex;

And the model is represented by the following equation:

-7.2036219	+ (0.015490665 * CPD)	–	[0.00001737645 * (CPD – 15)^3]	*for all values CPD>15*
		+	[0.00002192414 9 * (CPD – 20.185718)^3]	*for all values CPD>20*
		-	[0.0000045476985 * (CPD – 40)^3]	*for all values CDP>40*
	+ (0.020041889 * SMK)	+	[0.0000065443781 * (SMK – 27.6577)^3]	*for all values SMK>27*
		-	[0.000013947696 * (SMK – 40)^3]	*for all values SMK>40*
		+	[0.0000074033175 * (SMK-50.910335)^3]	*for all values SMK>50*
	- (0.023358962 * QUIT)	+	[0.0019208669 * QUIT)^3]	*for all values*
		-	[0.0020031611 * (QUIT – 0.50513347)^3]	*for all values QUIT>0*
		+	[0.000082294194 * (QUIT – 12.295688)^3]	*for all values QUIT>12*
	+ (0.099168033 * AGE)	+	[0.0000062174577 * (AGE – 53.459001)^3]	*for all values AGE>53*
		-	[0.000012115774 * (AGE – 61.954825)^3]	*for all values AGE>61*
		+	[0.0000058983164 * (AGE – 70.910335)^3]	*for all values AGE>70*
	+ [0.06084611] *if ASB=yes*			
	- [0.49042298] *if SEX=female*			

Note: Include all terms in the equation for which the requirements are met. For example, if an individual smokes 25 cigarettes per day, the value for the terms associated with CPD will be (0.015490665 * **25**) – [0.00001737645 * (**25** – 15)^3] + [0.000021924149 * (**25** – 20.185718)^3].

Figure 21: Bach Risk Model

The one-year probability of a diagnosis of lung cancer = $1-S_0^{e^{(model)}}$

Where,

S_0 = 0.99629;

CPD = cigarettes per day, SMK = duration of smoking, QUIT = duration of quitting, AGE = age, ASB = asbestos exposure, and SEX = sex;

And model is represented by the following equation:

-9.7960571	+ (0.060818386 * CPD)	–	[0.00014652216 * (CPD – 15)^3]	*for all values CPD>15*
		+	[0.00018486038 * (CPD 20.185718)^3]	*for all values CPD>20*
		-	[0.000033347226 * (CPD – 40)^2]	*for all values CPD>40*
	+ (0.11425297 * SMK)	-	[0.000080091477 * (SMK – 27.6577)^3]	*for all values SMK>27*
		+	[0.00017069483 * (SMK – 40)^3]	*for all values SMK>40*
		-	[0.000099603358 * (SMK-50.910335)^3]	*for all values SMK>50*
	- (0.085684793 * QUIT)	+	[0.0065499693 * QUIT)^3]	*for all values*
		-	[0.0068305845 * (QUIT – 0.50513347)^3]	*for all values QUIT>0*
		+	[0.00028061519 * (QUIT 12.295688)^2]	*for all values QUIT>12*
	+ (0.070322812 * AGE)	-	[0.00009382122 * (AGE – 53.459001)^3]	*for all values AGE>53*
		+	[0.00018282661 * (AGE – 61.954825)^3]	*for all values AGE>61*
		-	[0.000089005389 * (AGE – 70.910335)^3]	*for all values AGE>70*
	+ [0.2153936] *if ASB=yes*			
	- [0.05827261] *if SEX=female*			

Note: Include all terms in the equation for which the requirements are met. For example, if an individual smokes 25 cigarettes per day, the value for the terms associated with CPD will be (0.060818386 * **25**) – [0.00014652216 * (**25** – 15)^3] + [0.00018486938 * (**25** – 20.185718)^3].

Figure 22: Bach Risk Model

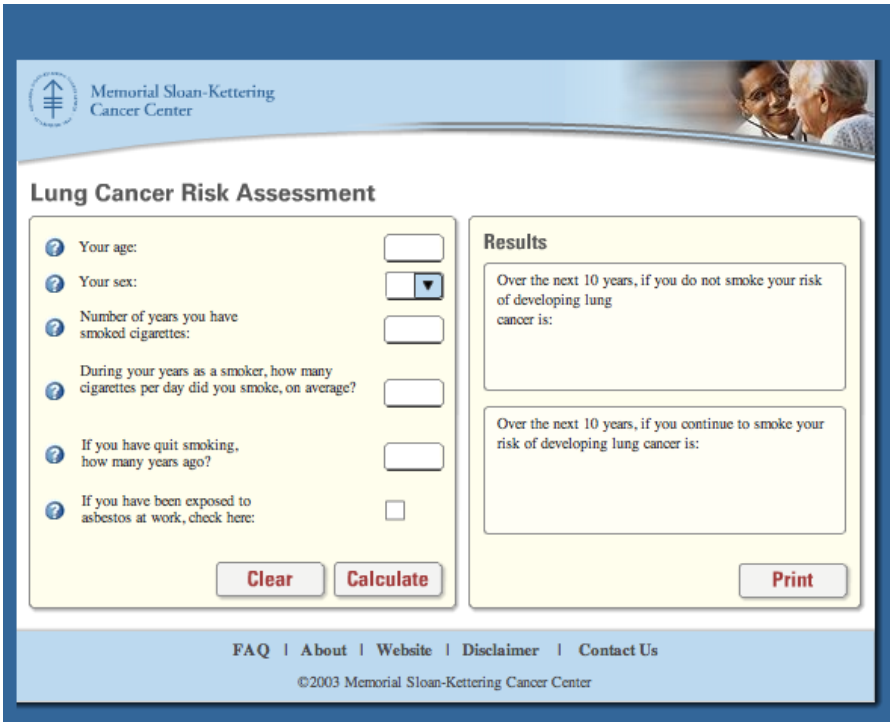

Figure 23: Online Ten Year Cancer Risk Assessment

	Male		Female	
Age group	Incidence rate[a]	α-value	Incidence rate[a]	α-value
40–44	15.5	−9.06	5.97	−9.90
45–49	37.87	−8.16	37.34	−8.06
50–54	88.65	−7.31	68.14	−7.46
55–59	172.26	−6.63	175.24	−6.50
60–64	329.02	−5.97	230.6	−6.22
65–69	487.42	−5.56	288.06	−5.99
70–74	616.45	−5.31	464.99	−5.49
75–79	950.61	−4.83	594.19	−5.23
80–84	1096.42	−4.68	497.09	−5.42

[a]Lung cancer incidence rate per 100 000 person-years, Liverpool, 2002–2004.

Figure 24: Age and sex specific lung cancer incidence rates and estimated values relating to 5-year absolute risk

Risk factor/category	Odds ratio[a]	(95% CI)	Odds ratio[b]	(95% CI)	P-value	Model coefficient
Smoking duration					<0.001	
Never	1.00	Reference	1.00	Reference		0.000
1–20 years	2.48	(1.47–4.17)	2.16	(1.21–3.85)		0.769
21–40 years	5.81	(3.68–9.18)	4.27	(2.62–6.94)		1.452
41–60 years	19.24	(12.07–30.67)	12.27	(7.41–20.30)		2.507
>60 years	41.74	(17.86–97.56)	15.25	(5.71–40.65)		2.724
Prior diagnosis of pneumonia					0.002	
No	1.00	Reference	1.00	Reference		0.000
Yes	1.62	(1.21–2.17)	1.83	(1.26–2.64)		0.602
Occupational exposure to asbestos				<0.001		
No	1.00	Reference	1.00	Reference		0.000
Yes	1.94	(1.46–2.59)	1.89	(1.35–2.62)		0.634
Prior diagnosis of malignant tumour					0.005	
No	1.00	Reference	1.00	Reference		0.000
Yes	2.55	(1.76–3.71)	1.96	(1.22–3.14)		0.675
Family history of lung cancer					0.01	
No	1.00	Reference	1.00	Reference		0.000
Early-onset (<60 years)	1.54	(1.03–2.29)	2.02	(1.18–3.45)		0.703
Late-onset (≥60 years)	1.08	(0.80–1.46)	1.18	(0.79–1.76)		0.168

[a]Odds ratios derived from univariate conditional logistic regression. [b]Odds ratios derived from multivariate conditional logistic regression.

Figure 25: LLP multivariate risk model, with unadjusted and adjusted odds ratios and 95% confidence intervals corresponding to the model coefficients

11.2 R Syntax

```
library(MASS)
library(Hmisc) #To run "Describe" Command
library(modeest) #To run "mlv" Command
res<-read.csv("Resol.csv") #Main sheet
attach(res)
health<-read.csv("HealthSheet.csv") #Health Sheet
attach(health)
resfh<-read.csv("Mainnhealth.csv") #Main sheet (Containing family history)
attach(resfh)
options(digits=4) #Number of digits to print on output
# Mean Age patients excluding missing values
a<-mean(AgeRegistered[!is.na(AgeRegistered)]);a
ncol(res) # Number of columns ReSoLuCENT main data sheet
nrow(res) # Number of rows of Health data sheet
ncol(health)
nrow(health)
# Mean Age Cases
b<-mean(AgeRegistered[CaseControl_c=="Case" ]);b
# Standard Error Age Cases
bvar<-var(AgeRegistered[CaseControl_c=="Case" ]);sqrt(bvar)
# Mean Age Controls
c<-mean(na.omit(AgeRegistered)[CaseControl_c=="Control" ]);c
# Standard Error Age Controls
cvar<-var(na.omit(AgeRegistered)[CaseControl_c=="Control" ]);sqrt(cvar)
hist(AgeRegistered[!is.na(AgeRegistered)],prob=TRUE,xlab="Age Registered",main="")
lines(density(AgeRegistered[!is.na(AgeRegistered)]))
describe(res)
# Barplot Case/Control freq.
plot(CaseControl_c,xlab="Subject",ylab="Frequency")
# Barplots Case - Control Gender Frequency
par(mfrow=c(1,2))
barplot(table(Sex,CaseControl_c),xlab="Subject",ylab="Frequency",
legend.text=c("Male","Female"), ylim= c(0,550))
barplot(table(Sex,CaseControl_c),xlab="Subject",ylab="Frequency",
```

```
legend.text=c("Male","Female"),beside=TRUE, ylim= c(0,300))
# Number of Male and Female Cases #1(male) and #2(Female)
sum(res$CaseControl_c=="Case" & res$Sex=="1", na.rm = TRUE)
sum(res$CaseControl_c=="Case" & res$Sex=="2", na.rm = TRUE)
# Number of Male and Female Controls
sum(res$CaseControl_c=="Control" & res$Sex=="1", na.rm = TRUE)
sum(res$CaseControl_c=="Control" & res$Sex=="2", na.rm = TRUE)
# Controls Alive
sum((Status =="Alive")[CaseControl_c=="Control" ])
# Cases Alive
sum((Status =="Alive")[CaseControl_c=="Case" ])
# Controls Dead
879-sum((Status =="Alive")[CaseControl_c=="Control" ])
# Controls Dead
879-sum((Status =="Alive")[CaseControl_c=="Case" ])
summary(S28)
#Cases Current-smoker
sum((S28 =="Yes")[CaseControl_c=="Case" ]) ;150/516
# Cases Ex-smoker
sum((S28 =="Ex")[CaseControl_c=="Case" ]) ;245/516
#Controls Current-Smoker
sum((S28 =="Yes")[CaseControl_c=="Control" ]);99/363
# Controls Ex-smoker
sum((S28 =="Ex")[CaseControl_c=="Control" ]); 100/363
#Dusty Conditions
describe(S19)
#Cases
sum((S19 =="Yes")[CaseControl_c=="Case" ])
sum((S19 =="No")[CaseControl_c=="Case" ])
#Controls
sum((S19 =="Yes")[CaseControl_c=="Control" ])
sum((S19 =="No")[CaseControl_c=="Control" ])
barplot(table(S19,CaseControl_c),xlab="Subject",ylab="Frequency",
legend.text=c("Unknown","No","Yes"), ylim= c(0,550))
#Quit Duration
```



```
quit<-AgeRegistered -S31;quit   # Age registered - Age quit = years quit
quit<-quit[!is.na(quit)] # Exclude NAs ie. Current and Non-smokers
quit[(quit<0)]<- (-quit[(quit<0)])
mean(quit)
quit[336];S31[336];AgeRegistered[336]  #-0.4,52,51.6  Quit after registering
quit[587];S31[587];AgeRegistered[587]  #-0.1,57,56.9
quit[432];S31[432];AgeRegistered[432]  #-0.3,56,55.7
quit[1];S31[1];AgeRegistered[1]
#1.4,44,45.4  Ex-smoker who quit before registering
quit[2];S31[2];AgeRegistered[2]  #NA
#Current smokers and Non-smokers
# Quitting duration for cases
casequit<-quit[CaseControl_c=="Case" ]
casequit <-casequit[!is.na(casequit)]
# Exclude NAs ie. Current and Non-smokers
mean(casequit)
# Quitting duration for controls
controlquit<-quit[CaseControl_c=="Control" ]
controlquit <-controlquit[!is.na(controlquit)]
# Exclude NAs ie. Current and Non-smokers
mean(controlquit)
par(mfrow=c(1,2))
hist(casequit,xlab="Case Quitting Duration",main="")
#Smoking Quitting frequency histogram
hist(controlquit,xlab="Control Quitting Duration",main="")
#Smoking Quitting frequenc histogram
#Smoking Duration
for(i in 1:length(S31))  #(Assigning 0 to NA values)
{if (is.na(S30[i])) {S30[i]=0}
{if (is.na(S31[i])) {S31[i]=0}}}
dur<-S31-S30
#Smoking Duration = Age Stopped smoking regularly - Age started
for(i in 1:length(dur))
{if (dur[i]<0) {dur[i]= AgeRegistered[i]-S30[i]}}
#Current Smokers smoking duration
```

```
#Assuming negative value (subject 431 with missing age) is 0 (ie non-smoker)
dur[is.na(dur)]<-0
dur<-dur[dur>0] #Exclude 0 values ie. Non-smokers
# To calculate overall mean and variance
mean(dur) #32.34
durcase<-dur[CaseControl_c=="Case" ];durcase # Smoking Duration for Cases
durcase<-durcase[!is.na(durcase)]
length(durcase)
mean(durcase)
mlv(durcase, method = "mfv")
durcont<-dur[CaseControl_c=="Control" ];durcont # Smoking Duration for Controls
durcont <-durcont[!is.na(durcont)]
length(durcont)
mean(durcont)
mlv(durcont, method = "mfv")
par(mfrow=c(1,2))
hist(durcase,xlab="Case Smoking Duration",main="")
#Smoking durion frequency histogram
hist(durcont,xlab="Control Smoking Duration",main="")
#Smoking durion frequenc histogram
# Cases, Ex-Smokers
Exdurcase<-durcase[durcase>0];Exdurcase
mean(Exdurcase)
var<-var(Exdurcase);sqrt(var)
# Controls, Ex-Smokers
Exdurcont<-durcont[durcont>0];Exdurcont
mean(Exdurcont)
var<-var(Exdurcont);sqrt(var)
#Run "dur<-S31-S30" , "durcase" and "durcont" again
# Current Case Smokers
Curdurcase<-rep(NA,length(durcase))
for(i in 1:length(durcase))
{if (durcase[i]<0) {Curdurcase[i]=AgeRegistered[i]-S30[i]}}
mean(Curdurcase[!is.na(Curdurcase)])
var<-var(Curdurcase[!is.na(Curdurcase)]) ;sqrt(var)
```



```
# Current Control Smokers
Curdurcont<-rep(NA,length(durcont))
for(i in 1:length(durcont))
{if (durcont[i]<0) {Curdurcont[i]=AgeRegistered[i]-S30[i]}}
mean(Curdurcont[!is.na(Curdurcont)])
var<-var(Curdurcont[!is.na(Curdurcont)]) ;sqrt(var)
# Mean and Standard Deviation Smoking per day at Age 20 #Cases
exsmokers<-S28[S28=="Ex"] # Ex-Smokers
currentsmokers<-S28[S28=="Yes"] # Current Smokers
SS31_20<-S31_20[S31_20>0]
case20<-SS31_20[CaseControl_c=="Case" ];case20
case20<-case20[!is.na(case20)]
case20ex<-case20[S28=="Ex"]  #Ex-Smokers
mean(case20ex[!is.na(case20ex)])
var<-var(case20ex[!is.na(case20ex)]);sqrt(var)
case20cur<-case20[S28=="Yes"]  # Current Smokers
mean(case20cur[!is.na(case20cur)])
var<-var(case20cur[!is.na(case20cur)]);sqrt(var)
#Controls
control20<-SS31_20[CaseControl_c=="Control" ];control20
control20ex<-control20[S28=="Ex"]  #Non-Smokers
mean(control20ex[!is.na(control20ex)])
var<-var(control20ex[!is.na(control20ex)]);sqrt(var)
# Current Smokers
control20cur<-control20[S28=="Yes"]  # Current Smokers
mean(control20cur[!is.na(control20cur)])
var<-var(control20cur[!is.na(control20cur)]);sqrt(var)
# Mean and Standard Deviation Smoking per day at Age 30
#Cases
SS31_30<-S31_30[S31_30>0]
case30<-SS31_30[CaseControl_c=="Case" ];case30
case30ex<-case30[S28=="Ex"]  #Ex-Smokers
mean(case30ex[!is.na(case30ex)])
var<-var(case30ex[!is.na(case30ex)]);sqrt(var)
case30cur<-case30[S28=="Yes"]  # Current Smokers
```

```
mean(case30cur[!is.na(case30cur)])
var<-var(case30cur[!is.na(case30cur)]);sqrt(var)
#Controls
control30<-SS31_30[CaseControl_c=="Control" ];control30
control30ex<-control30[S28=="Ex"]  #Ex-Smokers
mean(control30ex[!is.na(control30ex)])
var<-var(control30ex[!is.na(control30ex)]);sqrt(var)
control30cur<-control30[S28=="Yes"]  # Current Smokers
mean(control30cur[!is.na(control30cur)])
var<-var(control30cur[!is.na(control30cur)]);sqrt(var)
# Mean and Standard Deviation Smoking per day at Age 40
#Cases
SS31_40<-S31_40[S31_40>0]
case40<-SS31_40[CaseControl_c=="Case" ];case40
case40ex<-case40[S28=="Ex"]  #Ex-Smokers
mean(case40ex[!is.na(case40ex)])
var<-var(case40ex[!is.na(case40ex)]);sqrt(var)
case40cur<-case40[S28=="Yes"]  # Current Smokers
mean(case40cur[!is.na(case40cur)])
var<-var(case40cur[!is.na(case40cur)]);sqrt(var)
#Controls
control40<-SS31_40[CaseControl_c=="Control" ];control40
control40ex<-control40[S28=="Ex"]  #Ex-Smokers
mean(control40ex[!is.na(control40ex)])
var<-var(control40ex[!is.na(control40ex)]);sqrt(var)
control40cur<-control40[S28=="Yes"]  # Current Smokers
mean(control40cur[!is.na(control40cur)])
var<-var(control40cur[!is.na(control40cur)]);sqrt(var)
# Mean and Standard Deviation Smoking per day at Age 50
#Cases
SS31_50<-S31_50[S31_50>0]
case50<-SS31_50[CaseControl_c=="Case" ];case50
case50ex<-case50[S28=="Ex"]  #Ex-Smokers
mean(case50ex[!is.na(case50ex)])
var<-var(case50ex[!is.na(case50ex)]);sqrt(var)
```



```
case50cur<-case50[S28=="Yes"]  # Current Smokers
mean(case50cur[!is.na(case50cur)])
var<-var(case50cur[!is.na(case50cur)]);sqrt(var)
# Controls
control50<-SS31_50[CaseControl_c=="Control" ];control50
control50ex<-control50[S28=="Ex"]  #Ex-Smokers
mean(control50ex[!is.na(control50ex)])
var<-var(control50ex[!is.na(control50ex)]);sqrt(var)
control50cur<-control50[S28=="Yes"]  # Current Smokers
mean(control50cur[!is.na(control50cur)])
var<-var(control50cur[!is.na(control50cur)]);sqrt(var)
summary(S28)
#Treat blank cells as Unknown variable
S28[(S28!="Ex") & (S28!="No") & (S28!="Yes") ] <- NA
S28<-factor(S28,labels=c("Ex","No","Yes"))
# Regular smoker (10 cigarettes per week for a year)
barplot(table(CaseControl_c, S28),beside=TRUE
,xlab="Regular Smoker Subject",ylab="Frequency",legend.text=TRUE)
S29[S29==0]<-NA
# Age at first full cigarette
barplot(table(CaseControl_c, S29),xlab="Age",ylab="Frequency"
,beside=TRUE,legend.text=TRUE)
mean(S29[!is.na(S29)])
S30[S30==0]<-NA
# Starting age regular smoking
barplot(table(CaseControl_c, S30),xlab="Age",ylab="Frequency"
,beside=TRUE,legend.text=TRUE)
mean(S30[!is.na(S30)])
#Treat blank cells as Unknown variable
S37[(S37!="Yes") & (S37!="No") ] <- NA
S37<-factor(S37,labels=c("No","Yes"))
# Passive smoking
barplot(table( CaseControl_c,S37),
xlab="Passive Smoker Subjects",ylab="Frequency",
beside=TRUE,legend.text=TRUE)
```

```
#Treat blank cells as Unknown variable
S20[(S20!="No") & (S20!="Yes") ] <- NA
S20<-factor(S20,labels=c("No","Yes"))
# Asbestos Exposure
barplot(table(CaseControl_c,S20),
xlab="Asbestos Exposed Subjects",ylab="Frequency"
,beside=TRUE,legend.text=TRUE)
S31[S31==0]<-NA
# If stopped smoking, Age last smoked?
barplot(table(CaseControl_c,S31),xlab="Age",
ylab="Frequency",beside=TRUE,legend.text=TRUE)
# Mode smoking duraion age  # Overall
library(modeest)
mlv(S31[!is.na(S31)], method = "mfv")
# Mean smoking duraion age  # Overall
mean(S31[!is.na(S31)])
# Controls
controlce<-S31[CaseControl_c=="Control" ];controlce
mean(controlce[!is.na(controlce)])
# Cases
casece<-S31[CaseControl_c=="Case" ];casece
mean(casece[!is.na(casece)])
# Cigarette type at ages 20,30,40 and 50
par(mfrow=c(2,2))
barplot(table(CaseControl_c, S31_20Type),xlab="Cigarette Type at Age 20",
beside=TRUE,legend.text=TRUE)
barplot(table(CaseControl_c, S31_30Type),xlab="Cigarette Type at Age 30",
beside=TRUE,legend.text=TRUE)
barplot(table(CaseControl_c, S31_40Type),xlab="Cigarette Type at Age 40",
beside=TRUE,legend.text=TRUE)
barplot(table(CaseControl_c, S31_50Type),xlab="Cigarette Type at Age 50",
beside=TRUE,legend.text=TRUE)
# Assign "N/A" to large negative and positive values of Period Alive
for(i in 1:length(PeriodAlive))
{if (PeriodAlive[i]<1) {PeriodAlive[i]=NA}  #Assigns N/A to negatives and Nils
```



```
else
{ if (PeriodAlive[i]>90) {PeriodAlive[i]=NA} #Excludes very large values
else PeriodAlive[i]=PeriodAlive[i]}}
PeriodAlive<-PeriodAlive[!is.na(PeriodAlive)] #Excludes N/A s
length(PeriodAlive)  #209 alive after diagnosis
mean(PeriodAlive)  # 12.85 months Calculate average months alive after diagnosis
# Status
barplot(table(CaseControl_c,Status[!is.na(Status)]),legend.text=TRUE)
# Describe own health
barplot(table(CaseControl_c,S05[!is.na(S05)]),beside=TRUE,legend.text=TRUE)
# Occupation Pie Chart
lbls <- paste(names(table(S16)), "\n", table(S16), sep="")
pie(table(S16), labels = lbls,main="Occupation Pie Chart\n (with sample sizes)")
#Assign NA to missing values
Sitelung[(Sitelung!="Left lung") & (Sitelung!="Right lung") ] <- NA
Sitelung <-factor(Sitelung,labels=c("Left lung","Right lung"))
#Lung Site
barplot(table(Sitelung),ylim=c(0,280),xlab="Lung Site",ylab="Frequency")
# Sum of subjects living with partener/spouse
for(i in 1:length(S04_1[CaseControl_c=="Case" ]))
{if (is.na(S04_1[i])) {S04_1[i]=NA}
else
{ if (S04_1[i]>=1) {S04_1[i]=1}
else S04_1[i]=0}}
sum(S04_1[!is.na(S04_1)])
# Sum of subjects living with children/stepchildren
for(i in 1:length(S04_2))
{if (is.na(S04_2[i])) {S04_2[i]=NA}
else
{ if (S04_2[i]>=1) {S04_2[i]=1}
else S04_2[i]=0}}
sum(S04_2[!is.na(S04_2)])
# Sum of subjects living with Siblings
for(i in 1:length(S04_4))
{if (is.na(S04_4[i])) {S04_4[i]=NA}
```

```
else
{ if (S04_4[i]>=1) {S04_4[i]=1}
else S04_4[i]=0}}
sum(S04_4[!is.na(S04_4)])
# Health Spreadsheet Discriptive Satistics
describe(health)
describe(health$H1Alive) #Family Alive
summary(health$H1Cause) # Cause of Death
describe(health$H3) #Family Smoking Status
summary(health$HCa) #Has had Cancer
summary(health$H_2Ca) #More than 1 Cancer
summary(health$HCaSite1) #Part of Body Affected
summary(health$HCaSite2) #Cancer Site 2
#Family Hisotry on Health Sheet (health data)
FH
fh<-H1Relationship[FH=="1"]
onefh<-na.exclude(fh)
describe(fh) #142
describe(onefh)
summary(fh)
#(Mainnhealth) sheet
describe(Family) #134
caseFamily<-Family[CaseControl_c=="Case"]  #Cases
describe(caseFamily)
case1<-caseFamily[caseFamily=="1"];na.exclude(case1) #1 Family memebers
describe(case1)
case2<-caseFamily[caseFamily=="2"];na.exclude(case2) #2 Family memebers
describe(case2)
ControlFamily<-Family[CaseControl_c=="Control"]  #Controls
describe(ControlFamily)
control1<-ControlFamily[ControlFamily=="1"];na.exclude(control1)
#1 Family member
describe(control1)
control2<-ControlFamily[ControlFamily=="2"];na.exclude(control2)
#2 Family members
```



```
describe(control2)
caseover<-caseFamily[AgeRegistered>60] # Over 60 Family memeber Cases
describe(caseover)
controlover<-ControlFamily[AgeRegistered>60] # Over 60 Family memeber Control
describe(controlover)
casebelow<-caseFamily[AgeRegistered<60]
# Below 60 Family memeber Cases
describe(casebelow)
controlbelow<-ControlFamily[AgeRegistered<60] # Below 60 Family memeber Control
describe(controlbelow)
##### Pack Years#####
# Assign "Nil" to unknown values
for(i in 1:length(S31))
{if (is.na(S30[i])) {S30[i]=0} #Non-Smokers
{if (is.na(S31[i])) {S31[i]=0}           #Current Smokers
{if (is.na(AgeRegistered[i])) {AgeRegistered[i]=0}   #Age joined the study
{if (is.na(S31_20[i])) {S31_20[i]=0}    #Non smoker at age 20
{if (is.na(S31_30[i])) {S31_30[i]=0}    #Non smoker at age 30
{if (is.na(S31_40[i])) {S31_40[i]=0}    #Non smoker at age 40
{if (is.na(S31_50[i])) {S31_50[i]=0}    #Non smoker at age 50
}}}}}}}
# Pack Years
p<-rep(NA,length(S31)) #Create an array of NAs with length S31
for(i in 1:length(S31))
{if (S30[i]==0 && S31[i]==0) {p[i]=0}   # Non-smokers
else
{if ((S31[i]==0) && AgeRegistered[i]>=50)
# Current smoker ,current age over 50
{p[i]<-((AgeRegistered[i]-50)*S31_50[i]+10*S31_40[i]
+10*S31_30[i]+10*S31_20[i]+((20-S30[i])*S31_20[i]))/20}
else
{if ((S31[i]==0) && AgeRegistered[i]>=40)
# Current smoker ,current age over 40
{p[i]<-((AgeRegistered[i]-40)*S31_40[i]+10*S31_30[i]
+10*S31_20[i]+((20-S30[i])*S31_20[i]))/20}
```

```
else
{if ((S31[i]==0) && AgeRegistered[i]>=30)    # Current smoker ,current age over 30
{p[i]<-((AgeRegistered[i]-30)*S31_30[i]+10*S31_20[i]
+((20-S30[i])*S31_20[i]))/20}
else
{if ((S31[i]==0) && AgeRegistered[i]>=20)   # Current smoker ,current age over 20
{p[i]<-((AgeRegistered[i]-20)*S31_20[i]+((20-S30[i])*S31_20[i]))/20} else
{if ((S31[i]==0) && AgeRegistered[i]>=15)  # Current smoker ,current age over 15
{p[i]<-((AgeRegistered[i]-S30[i])*S31_20[i])/20}
else
{if (S31[i]>=50)   # Stopped smoking after 50
{p[i]<-(((S31[i]-50)*S31_50[i])+10*S31_40[i]+10*S31_30[i]
+10*S31_20[i]+((20-S30[i])*S31_20[i]))/20}
else
{if (S31[i]>=40)  #  Stopped smoking after 40
{p[i]<-(((S31[i]-40)*S31_40[i])+10*S31_30[i]+10*S31_20[i]+((20-S30[i])*S31_20[i]))/20}
else
{if (S31[i]>=30)  #  Stopped smoking after 30
{p[i]<-(((S31[i]-30)*S31_30[i])+10*S31_20[i]+((20-S30[i])*S31_20[i]))/20}
else
{if (S31[i]>=20)  # Stopped smoking after 20
{p[i]<-((S31[i]-20)*S31_20[i]+((20-S30[i])*S31_20[i]))/20}
else
{if (S31[i]>=17)
{p[i]<-((S31[i]-S30[i])*S31_20[i])/20}  # Stopped smoking after 17
}}}}}}}}}}
# Include starting smoking ages in bands 20-30, 30-40
and 40-50 for current and ex-smokers both
for(i in 1:length(S31))
{if (p[i]>0 && 30 >S30[i] && S30[i]>20) {p[i]<-p[i]+((30-S30[i])*S31_30[i])/20}
else
{if (p[i]>0 && 40 >S30[i] && S30[i]>30) {p[i]<-p[i]+((40-S30[i])*S31_40[i])/20}
else
{if (p[i]>0 && 50 >S30[i] && S30[i]>40) {p[i]<-p[i]+((50-S30[i])*S31_50[i])/20}}}}
meanp<-mean(p[!is.na(p)]);meanp #mean p  21.18
```



```
# Diagram
#Pack Years Case Controls
max(p[!is.na(p)])
median(p[!is.na(p)])
median(packcase)
median(packcontrol)
# Mean and Standard Deviation Pack Years
#Cases
packcase<-p[CaseControl_c=="Case" ];packcase   #Pack Years for relevant Cases
packcase<-packcase[!is.na(packcase)]  #Exclude NA values
mean(packcase)
var<-var(packcase);sqrt(var)
#Controls
packcontrol<-p[CaseControl_c=="Control" ];packcontrol
#Pack Years for relevant Controls
packcontrol<-packcontrol[!is.na(packcontrol)] #Exclude NA values
mean(packcontrol)
var<-var(packcontrol);sqrt(var)
# Pack Year Box-plot
boxplot(p,packcontrol, packcase,notch=TRUE,
names=c("All Subjects","Controls","Cases"),ylab="Pack Years")
# Pack Years for Cases (Ex-Smokers and Current Smokers)
exsmokers<-S28[S28=="Ex"] # Ex-Smokers
currentsmokers<-S28[S28=="Yes"] # Current Smokers
currentpckcase<-packcase[S28=="Yes"]
currentpckcase <-currentpckcase[!is.na(currentpckcase)]
# Current Smokers Cases Pack Years
mean(currentpckcase)
var<-var(currentpckcase);sqrt(var)
Expackcase<-packcase[S28=="Ex"]
Expackcase<-Expackcase[!is.na(Expackcase)] # Ex-smoker Cases Pack Years
mean(Expackcase)
var<-var(Expackcase);sqrt(var)
# Pack Years for Controls (Ex-Smokers and Current Smokers)
Currentpackcontrol<-packcontrol[S28=="Yes"]
```

```
# Current Smokers controls pack years
Currentpackcontrol <-Currentpackcontrol[!is.na(Currentpackcontrol)]
mean(Currentpackcontrol)
var<-var(Currentpackcontrol);sqrt(var)
Expackcontrol<-packcontrol[S28=="Ex"] # Ex-smoker controls pack years
Expackcontrol<-Expackcontrol[!is.na(Expackcontrol)]
mean(Expackcontrol)
var<-var(Expackcontrol);sqrt(var)
#### Bach Risk Model ####
###Smoking Duration###
for(i in 1:length(S31))  #(Assigning 0 to NA values)
{if (is.na(S30[i])) {S30[i]=0}
{if (is.na(S31[i])) {S31[i]=0}}}
dur<-S31-S30  #Smoking Duration (Ex-smokers) = Age Stopped smoking regularly - Age started
for(i in 1:length(dur))
{if (dur[i]<0) {dur[i]= AgeRegistered[i]-S30[i]}}  #Current Smokers smoking duration
#Assuming negative value (subject 431 with missing age) is 0 (ie non-smoker)
dur[is.na(dur)]<-0
#Quit Duration
quit<-AgeRegistered -S31;quit   # Age registered - Age quit = years quit
for(i in 1:879)
{if (S31[i]==0) {quit[i]=0} #Assing "NA" to nagtive values (current smokers)
{if (quit[i]<0) {quit[i]=0}}} #Assing 0 to nagtive values (S31>Age)
# As cigarettes per day are used in the calculation of pack years,
I take account the
# starting age and stopping age to calculate the weighted average
#Only for purpose of calculating the risk model:
To assing NAs (subject 835 and 431) "Nil" value
for(i in 1:length(p))
{if (is.na(p[i])) {p[i]=0}}  #Pack years
CPD<-rep(NA,length(p)) #Create an array of NAs with length S31
for(i in 1:length(p))
{if (S30[i]==0 && S31[i]==0) {CPD[i]=0} # Non-smokers
else
{if ((S31[i]==0)) {CPD[i]<-20*p[i]/(AgeRegistered[i]-S30[i])}  # Current smoker
```



```
else
{CPD[i]<-20*p[i]/(S31[i]-S30[i])}}}  # Ex-smoker
mean(CPD)  #12.42
# oneyeardiagprob :  one-year probability of a diagnosis of lung cancer
# oneyeardeathprob:one-year probability of death
in the absence of a diagnosis of lung cancer
####durone, durtwo, durthree, durfour######
#  CPD dur+1(if current smoker) quit AgeRegistered+1 S20 Sex
durone<-rep(NA,length(dur))
durtwo<-rep(NA,length(dur))
durthree<-rep(NA,length(dur))
durfour<-rep(NA,length(dur))
for (i in 1:length(dur))    # If current smoker year add up
{if  (S31[i]=="0") {durone[i]<-dur[i]+1}
 #Duration of smoking afre one year to calculate 2 years risk
{if  (S31[i]=="0") {durtwo[i]<-dur[i]+2}
{if  (S31[i]=="0") {durthree[i]<-dur[i]+3}
{if  (S31[i]=="0") {durfour[i]<-dur[i]+4}
# Non-smoker
{if  (dur[i]=="0") {durone[i]<-0}
{if  (dur[i]=="0") {durtwo[i]<-0}
{if  (dur[i]=="0") {durthree[i]<-0}
{if  (dur[i]=="0") {durfour[i]<-0}
# Ex-smoker
{if  (S31[i]>"0") {durone[i]<-dur[i]}
{if  (S31[i]>"0") {durtwo[i]<-dur[i]}
{if  (S31[i]>"0") {durthree[i]<-dur[i]}
{if  (S31[i]>"0") {durfour[i]<-dur[i]}
}}}}}}}}}}}
##One-year probability of death in the absence of diagnosis of lung cancer##
#### CPD ####
model1d<-rep(NA,879) #Create an array of NAs
for(i in 1:length(CPD))
{if (CPD[i]>15) {model1d[i]<-(0.015490665*CPD[i])
-7.2036219-(0.00001737645 * ((CPD[i]-15)^3))}
```

```
{if (CPD[i]>20) {model1d[i]<-model1d[i]
+(0.000021924149 *(CPD[i]-20.185718)^3)}
{if (CPD[i]>40 && CPD[i]<60) {model1d[i]<-model1d[i]
-(0.0000045476985 * (CPD[i]-40)^3)}
}}} #### Duration: SMK (dur)####
model2d<-rep(NA,879) #Create an array of NAs
for(i in 1:length(dur))
{if (dur[i]>27) {model2d[i]<-(0.020041889 * dur[i])
+0.0000065443781 * ((dur[i]-27.6577) ^3)}
{if (dur[i]>40) {model2d[i]<- model2d[i]-(0.000013947696 * (dur[i] -40)^ 3) }
{if (dur[i]>50  && dur[i]<50) {model2d[i]<- model2d[i]
+(0.0000074033175 * (dur[i]-50.910335)^ 3) }
}}}
model2done<-rep(NA,879) #Create an array of NAs
for(i in 1:length(durone))
{if (durone[i]>27) {model2done[i]<-(0.020041889 * durone[i])
 +0.0000065443781 * ((durone[i]-27.6577) ^3)}
{if (durone[i]>40) {model2done[i]<- model2done[i]
-(0.000013947696 * (durone[i] -40)^ 3) }
{if (durone[i]>50  && durone[i]<50) {model2done[i]<-
model2d[i]+(0.0000074033175 * (durone[i]-50.910335)^ 3) }
}}}
model2dtwo<-rep(NA,879) #Create an array of NAs
for(i in 1:length(durtwo))
{if (durtwo[i]>27) {model2dtwo[i]<-(0.020041889 * durtwo[i])
+0.0000065443781 * ((durtwo[i]-27.6577) ^3)}
{if (durtwo[i]>40) {model2dtwo[i]<- model2dtwo[i]
-(0.000013947696 * (durtwo[i] -40)^ 3) }
{if (durtwo[i]>50  && durtwo[i]<50) {model2dtwo[i]<-
model2dtwo[i]+(0.0000074033175 * (durtwo[i]-50.910335)^ 3) }
}}}
model2dthree<-rep(NA,879) #Create an array of NAs
for(i in 1:length(durthree))
{if (durthree[i]>27) {model2dthree[i]<-(0.020041889 * durthree[i])
+0.0000065443781 * ((durthree[i]-27.6577) ^3)}
```



```
{if (durthree[i]>40) {model2dthree[i]<- model2dthree[i]
-(0.000013947696 * (durthree[i] -40)^ 3) }
{if (durthree[i]>50  && durthree[i]<50) {model2dthree[i]
<- model2dthree[i]+(0.0000074033175 * (durthree[i]-50.910335)^ 3) }
}}}
model2dfour<-rep(NA,879) #Create an array of NAs
for(i in 1:length(durfour))
{if (durfour[i]>27) {model2dfour[i]<-(0.020041889 * durfour[i])
+0.0000065443781 * ((durfour[i]-27.6577) ^3)}
{if (durfour[i]>40) {model2dfour[i]<- model2dfour[i]
-(0.000013947696 * (durfour[i] -40)^ 3) }
{if (durfour[i]>50  && durfour[i]<50) {model2dfour[i]<-
model2dfour[i]+(0.0000074033175 * (durfour[i]-50.910335)^ 3) }
}}} #### Quitting Duration (QUIT)####
model3d<-rep(NA,879) #Create an array of NAs
for(i in 1:length(quit))
{model3d[i]<-(0.0019208669*(quit[i]^3))-(0.023358962 * quit[i])}
{if (quit[i]>0) {model3d[i]<-model3d[i]- (0.0020031611 * (quit[i]- 0.50513347) ^3)
{if (quit[i]>12 && quit[i]<20) {model3d[i]<-model3d[i]
+(0.000082294194 * (quit[i] -12.295688) ^3)}
}} #### Age (AgeRegistered)####
model4d<-rep(NA,879) #Create an array of NAs with length S31
for(i in 1:length(AgeRegistered))
{if (AgeRegistered[i]>53)
{model4d[i]<-(0.099168033*AgeRegistered[i])+
(0.0000062174577*((AgeRegistered[i]-53.459001)^3))}
{if (AgeRegistered[i]>61)
{model4d[i]<-model4d[i]-0.000012115774*((AgeRegistered[i]-61.954825)^3)}
{if (AgeRegistered[i]>70 && AgeRegistered[i]<75)
{model4d[i]<-model4d[i]+(0.0000058983164 *((AgeRegistered[i]-70.910335)^3) )}
}}}
model4done<-rep(NA,879) #Create an array of NAs with length S31
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+1)>53)
{model4done[i]<-(0.099168033*(AgeRegistered[i]+1))
```

```
+(0.0000062174577*(((AgeRegistered[i]+1)-53.459001)^3))}
{if ((AgeRegistered[i]+1)>61)
{model4done[i]<-model4done[i]-0.000012115774
*(((AgeRegistered[i]+1)-61.954825)^3)}
{if ((AgeRegistered[i]+1)>70 && (AgeRegistered[i]+1)<75)
{model4done[i]<-model4done[i]+(0.0000058983164
*(((AgeRegistered[i]+1)-70.910335)^3) )}
}}}
model4dtwo<-rep(NA,879) #Create an array of NAs with length S31
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+2)>53)
{model4dtwo[i]<-(0.099168033*(AgeRegistered[i]+2))
+(0.0000062174577*(((AgeRegistered[i]+2)-53.459001)^3))}
{if ((AgeRegistered[i]+2)>61)
{model4dtwo[i]<-model4dtwo[i]-0.000012115774
*(((AgeRegistered[i]+2)-61.954825)^3)}
{if ((AgeRegistered[i]+2)>70 && AgeRegistered[i]<75)
{model4dtwo[i]<-model4dtwo[i]+(0.0000058983164
*(((AgeRegistered[i]+2)-70.910335)^3) )}
}}}
model4dthree<-rep(NA,879) #Create an array of NAs with length S31
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+3)>53)
{model4dthree[i]<-(0.099168033*(AgeRegistered[i]+3))
+(0.0000062174577*(((AgeRegistered[i]+3)-53.459001)^3))}
{if ((AgeRegistered[i]+3)>61)
{model4dthree[i]<-model4dthree[i]-0.000012115774
*(((AgeRegistered[i]+3)-61.954825)^3)}
{if ((AgeRegistered[i]+3)>70 && (AgeRegistered[i]+3)<75)
{model4dthree[i]<-model4dthree[i]+(0.0000058983164
 *(((AgeRegistered[i]+3)-70.910335)^3) )}
}}}
model4dfour<-rep(NA,879) #Create an array of NAs with length S31
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+4)>53)
```



```
{model4dfour[i]<-(0.099168033*(AgeRegistered[i]+4))
+(0.0000062174577*(((AgeRegistered[i]+4)-53.459001)^3))}
{if ((AgeRegistered[i]+4)>61)
{model4dfour[i]<-model4dfour[i]-0.000012115774
*(((AgeRegistered[i]+4)-61.954825)^3)}
{if ((AgeRegistered[i]+4)>70 && AgeRegistered[i]<75)
{model4dfour[i]<-model4dfour[i]+(0.0000058983164
*(((AgeRegistered[i]+4)-70.910335)^3) )}
}}} #### ASB = asbestos exposure (S20)####
model5d<-rep(NA,879) #Create an array of NAs
for(i in 1:879)
{if (S20[i]=="Yes") {model5d[i]<-0.06084611}
else   {model5d[i]<-0} }
#### Sex ####
for(i in 1:length(Sex))
{if (is.na(Sex[i])) {Sex[i]=1}}
model6d<-rep(NA,879) #Create an array of NAs
for (i in 1:879)
{if (Sex[i]=="2") {model6d[i]=(-0.49042298)}
{if (Sex[i]=="1") {model6d[i]=0}
}} #### Model ####
modeld<-rep(NA,879) #Create an array of NAs
modeldone<-rep(NA,879) #Create an array of NAs
modeldtwo<-rep(NA,879) #Create an array of NAs
modeldthree<-rep(NA,879) #Create an array of NAs
modeldfour<-rep(NA,879) #Create an array of NAs
for (i in 1:879){
{modeld[i]<-model1d[i] + model2d[i] + model3d[i] +
model4d[i]+ model5d[i]+ model6d[i]}
{modeldone[i]<-model1d[i] + model2done[i] + model3d[i] +
model4done[i]+ model5d[i]+ model6d[i]}
{modeldtwo[i]<-model1d[i] + model2dtwo[i] + model3d[i]
+model4dtwo[i]+ model5d[i]+ model6d[i]}
{modeldthree[i]<-model1d[i] + model2dthree[i] + model3d[i]
+model4dthree[i]+ model5d[i]+ model6d[i]}
```

```
{modeldfour[i]<-model1d[i] + model2dfour[i] + model3d[i]
+model4dfour[i]+ model5d[i]+ model6d[i]}}
# One-year probability of a death in the absence of a diagnosis of lung cancer
# = 1 - (S_0)*(e^{model}) for S_0 =1 - (0.9917663)^(e^{model})
oneyeardeathprob<-rep(NA,length(modeld)) #Create an array of NAs with length S3
twoyeardeathprob<-rep(NA,length(modeld)) #Create an array of NAs with length S31
threeyeardeathprob<-rep(NA,length(modeld)) #Create an array of NAs with length S31
fouryeardeathprob<-rep(NA,length(modeld)) #Create an array of NAs with length S31
fiveyeardeathprob<-rep(NA,length(modeld)) #Create an array of NAs with length S31
for(i in 1:length(modeld)){
{oneyeardeathprob[i]<- 1-(0.9917663^(exp(-modeld[i])))}
{twoyeardeathprob[i]<- 1-(0.9917663^(exp(-modeldone[i])))}
{threeyeardeathprob[i]<- 1-(0.9917663^(exp(-modeldtwo[i])))}
{fouryeardeathprob[i]<- 1-(0.9917663^(exp(-modeldthree[i])))}
{fiveyeardeathprob[i]<- 1-(0.9917663^(exp(-modeldfour[i])))}}
#####One-year probability of a diagnosis of lung cancer######
#### CPD ####
model1<-rep(NA,879) #Create an array of NAs
for(i in 1:length(CPD))
{if (CPD[i]>15) {model1[i]<-(0.060818386*CPD[i])-9.7960571-(0.00014652216 * ((CPD[i]-15)^3))}
{if (CPD[i]>20) {model1[i]<-model1[i]+(0.00018486938*(CPD[i]-20.185718)^3)}
{if (CPD[i]>40 && CPD[i]<60) {model1[i]<-model1[i]-(0.000038347226*(CPD[i]-40)^3)}
}}}
#### Duration: SMK (dur)####
model2<-rep(NA,879) #Create an array of NAs
for(i in 1:length(dur))
{if (dur[i]>27) {model2[i]<-(0.11425297*dur[i])-(0.000080091477 * (dur[i]-27.6577) ^3)}
{if (dur[i]>40) {model2[i]<-model2[i]+(0.00017069483 * (dur[i] -40)^ 3) }
{if (dur[i]>50 && dur[i]<50) {model2[i]<-model2[i]-
(0.000090603358 * (dur[i]-50.910335)^ 3) }
}}}
model2one<-rep(NA,879) #Create an array of NAs
for(i in 1:length(dur))
{if (durone[i]>27) {model2one[i]<-(0.11425297*durone[i])
-(0.000080091477 * (durone[i]-27.6577) ^3)}
```



```
{if (durone[i]>40) {model2one[i]<-model2one[i]
+(0.00017069483 * (durone[i] -40)^ 3) }
{if (durone[i]>50 && durone[i]<50) {model2one[i]<-model2one[i]
-(0.000090603358 * (durone[i]-50.910335)^ 3) }
}}}
model2two<-rep(NA,879) #Create an array of NAs
for(i in 1:length(dur))
{if (durtwo[i]>27) {model2two[i]<-(0.11425297* durtwo[i])
-(0.000080091477 * (durtwo[i]-27.6577) ^3)}
{if (durtwo[i]>40) {model2two[i]<-model2two[i]
+(0.00017069483 * (durtwo[i] -40)^ 3) }
{if (durtwo[i]>50 && durtwo[i]<50) {model2two[i]<-model2two[i]
-(0.000090603358 * (durtwo[i]-50.910335)^ 3) }
}}}
model2three<-rep(NA,879) #Create an array of NAs
for(i in 1:length(durthree))
{if (durthree[i]>27) {model2three[i]<-(0.11425297* durthree[i])
-(0.000080091477 * (durthree[i]-27.6577) ^3)}
{if (durthree[i]>40) {model2three[i]<-model2three[i]
+(0.00017069483 * (durthree[i] -40)^ 3) }
{if (durthree[i]>50 && durthree[i]<50) {model2three[i]<-
model2three[i]-(0.000090603358 * (durthree[i]-50.910335)^ 3) }
}}}
model2four<-rep(NA,879) #Create an array of NAs
for(i in 1:length(durfour))
{if (durfour[i]>27) {model2four[i]<-(0.11425297* durfour[i])
-(0.000080091477 * (durfour[i]-27.6577) ^3)}
{if (durfour[i]>40) {model2four[i]<-model2four[i]+
(0.00017069483 * (durfour[i] -40)^ 3) }
{if (durfour[i]>50 && durfour[i]<50) {model2four[i]<-
model2four[i]-(0.000090603358 * (durfour[i]-50.910335)^ 3) }
}}} #### Quitting Duration (QUIT)####
model3<-rep(NA,879) #Create an array of NAs
for(i in 1:length(quit))
{model3[i]<-(0.0065499693*(quit[i] ^3))-(0.085684793*quit[i])}
```

```
{if (quit[i]>0) {model3[i]<-model3[i]-(0.0068305845*(quit[i]-0.50513347) ^3)}
{if (quit[i]>12 && quit[i]<20) {model3[i]<-model3[i]+(0.00028061519 * (quit[i]-12.295688) ^3)}
}}
#### Age (AgeRegistered)####
model4<-rep(NA,879) #Create an array of NAs
for(i in 1:length(AgeRegistered))
{if (AgeRegistered[i]>53) {model4[i]<-(0.070322812*AgeRegistered[i])-
(0.00009382122*(AgeRegistered[i]-53.459001)^3)}
{if (AgeRegistered[i]>61) {model4[i]<-model4[i]
+(0.00018282661*(AgeRegistered[i]-61.954825)^3)}
{if (AgeRegistered[i]>70  && AgeRegistered[i]<75)
{model4[i]<-model4[i]-(0.000089005389 *(AgeRegistered[i]-70.910335)^3)}
}}}
model4one<-rep(NA,879) #Create an array of NAs
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+1)>53) {model4one[i]<-(0.070322812*(AgeRegistered[i]+1))-
(0.00009382122*(AgeRegistered[i]-53.459001)^3)}
{if ((AgeRegistered[i]+1)>61)
 {model4one[i]<-model4one[i]+(0.00018282661
*((AgeRegistered[i]+1)-61.954825)^3)}
{if ((AgeRegistered[i]+1)>70  && (AgeRegistered[i]+1)<75)
{model4one[i]<-model4one[i]-(0.000089005389
*((AgeRegistered[i]+1)-70.910335)^3)}}}}
model4two<-rep(NA,879) #Create an array of NAs
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+2)>53) {model4two[i]<-(0.070322812*(AgeRegistered[i]+2))
-(0.00009382122*((AgeRegistered[i]+2)-53.459001)^3)}
{if ((AgeRegistered[i]+2)>61) {model4two[i]<-model4two[i]+(0.00018282661
*((AgeRegistered[i]+2)-61.954825)^3)}
{if ((AgeRegistered[i]+2)>70  && (AgeRegistered[i]+2)<75)
{model4two[i]<-model4two[i]-(0.000089005389 *((AgeRegistered[i]+2)-70.910335)^3)}
}}}
model4three<-rep(NA,879) #Create an array of NAs
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+3)>53)
```



```
{model4three[i]<-(0.070322812*(AgeRegistered[i]+3))
-(0.00009382122*(AgeRegistered[i]-53.459001)^3)}
{if ((AgeRegistered[i]+3)>61)
{model4three[i]<-model4three[i]+(0.00018282661
*((AgeRegistered[i]+3)-61.954825)^3)}
{if ((AgeRegistered[i]+3)>70  && (AgeRegistered[i]+3)<75)
{model4three[i]<-model4three[i]-(0.000089005389
*((AgeRegistered[i]+3)-70.910335)^3)}}}}
model4four<-rep(NA,879) #Create an array of NAs
for(i in 1:length(AgeRegistered))
{if ((AgeRegistered[i]+4)>53)
{model4four[i]<-(0.070322812*(AgeRegistered[i]+4))
-(0.00009382122*(AgeRegistered[i]-53.459001)^3)}
{if ((AgeRegistered[i]+4)>61)
{model4four[i]<-model4four[i]+(0.00018282661
*((AgeRegistered[i]+4)-61.954825)^3)}
{if ((AgeRegistered[i]+4)>70  && (AgeRegistered[i]+4)<75)
{model4four[i]<-model4four[i]-(0.000089005389
 *((AgeRegistered[i]+4)-70.910335)^3)}}}} #### ASB = asbestos exposure (S20)####
model5<-rep(NA,879) #Create an array of NAs
for(i in 1:length(S20))
{if (S20[i]=="Yes") {model5[i]<-0.2153936}
else   {model5[i]<-0} #### Sex ####
for(i in 1:length(Sex))
{if (is.na(Sex[i])) {Sex[i]=1}}
model6<-rep(NA,879) #Create an array of NAs
for (i in 1:879)
{if (Sex[i]=="2") {model6[i]=(-0.05827261)}
{if (Sex[i]=="1") {model6[i]=0}
}} #### Model ####
diagmodel<-rep(NA,879) #Create an array of NAs
diagmodelone<-rep(NA,879)
diagmodeltwo<-rep(NA,879)
diagmodelthree<-rep(NA,879)
diagmodelfour<-rep(NA,879)
```

```
for (i in 1:879){
{diagmodel[i]<-model1[i] + model2[i] + model3[i] +model4[i]+ model5[i]+ model6[i]}
{diagmodelone[i]<-model1[i] + model2one[i] + model3[i] + model4one[i]+ model5[i]+ model6[i]}
{diagmodeltwo[i]<-model1[i] + model2two[i] + model3[i] + model4two[i]+ model5[i]+ model6[i]}
{diagmodelthree[i]<-model1[i] + model2three[i] + model3[i] +
 model4three[i]+ model5[i]+ model6[i]}
{diagmodelfour[i]<-model1[i] + model2four[i] + model3[i]
+ model4four[i]+ model5[i]+ model6[i]}}
# The one-year probability of a diagnosis of lung cancer = 1 - (0.99629)*(e^{model})
#Create an array of NAs with length S31
oneyeardiagprob<-rep(NA,length(diagmodel))  #Probability of diagnosis 1year
twoyeardiagprob<-rep(NA,length(diagmodel)) #Probability of diagnosis in 2 years
threeyeardiagprob<-rep(NA,length(diagmodel)) #Probability of diagnosis in 3 years
fouryeardiagprob<-rep(NA,length(diagmodel)) #Probability of diagnosis in 4 years
fiveyeardiagprob<-rep(NA,length(diagmodel)) #Probability of diagnosis in 5 years
for(i in 1:length(diagmodel)){
{oneyeardiagprob[i]<-1-(0.99629^(exp(-diagmodel[i])))}
{twoyeardiagprob[i]<-1-(0.99629^(exp(-diagmodelone[i])))}
{threeyeardiagprob[i]<-1-(0.99629^(exp(-diagmodeltwo[i])))}
{fouryeardiagprob[i]<-1-(0.99629^(exp(-diagmodelthree[i])))}
{fiveyeardiagprob[i]<-1-(0.99629^(exp(-diagmodelfour[i])))}}
      #########
# oneyeardiagprob :  one-year probability of a diagnosis of lung cancer
# oneyeardeathprob:one-year probability of death in the absence of a diagnosis of lung cancer
#  5 year risk=
#  +(probability of a diagnosis of lung cancer at the end of year 1)
#  +(probability of Survival and no diagnosis
of lung cancer at the end of year 1 *
 probability of a diagnosis of lung cancer at the end of year 2)
#  +(probability of Survival and no diagnosis
of lung cancer at the end of year 1 and 2 *
probability of a diagnosis of lung cancer at the end of year 3)
#  +(probability of Survival and no diagnosis
of lung cancer at the end of year 1,2 and 3 *
probability of a diagnosis of lung cancer at the end of year 4)
```



```
#  +(probability of Survival and no diagnosis
of lung cancer at the end of year 1,2,3 and 4 *
probability of a diagnosis of lung cancer at the end of year 5)
Risk<-rep(NA,879)
for (i in 1:879){
Risk[i]<-oneyeardiagprob[i]*(1-oneyeardeathprob[i])
+ (1-oneyeardeathprob[i])*(1-twoyeardeathprob[i])*(1-oneyeardiagprob[i])
*twoyeardiagprob[i]+(1-oneyeardeathprob[i])*(1-twoyeardeathprob[i])
*(1-threeyeardeathprob[i])*(1-oneyeardiagprob[i])*(1-twoyeardiagprob[i])
*threeyeardiagprob[i]+(1-oneyeardeathprob[i])*(1-twoyeardeathprob[i])
*(1-threeyeardeathprob[i])*(1-fouryeardeathprob[i])*(1-oneyeardiagprob[i])
*(1-twoyeardiagprob[i])*(1-threeyeardiagprob[i]) *fouryeardiagprob[i]
+ (1-oneyeardeathprob[i])*(1-twoyeardeathprob[i])*(1-threeyeardeathprob[i])
*(1-fouryeardeathprob[i])*(1-fiveyeardeathprob[i])*(1-oneyeardiagprob[i])
*(1-twoyeardiagprob[i])*(1-threeyeardiagprob[i])
*(1-fouryeardiagprob[i])*fiveyeardiagprob[i]}
summary(na.exclude(Risk))
# Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
# 0.000000 0.008056 0.013650 0.015490 0.019550 0.073740
#Validation: http://jnci.oxfordjournals.org/content/95/6/470/suppl/DC1
######## Bach ROC Vurve  ############
#Exclude unknown values of Case Control with respect to uknown values of Risk
for(i in 1:879)
{if (is.na(Risk[i])) {CaseControl_c[i]=NA}}
Risk<-(na.exclude(Risk))
CaseControl_c<-(na.exclude(CaseControl_c))
#for(i in 1:length(alpha))
#{if (is.na(alpha[i])) {alpha[i]=0}}
CaseControl_c <-ifelse(CaseControl_c=="Case", 1, 0)
#Recode to Case=1 and Control=0
length(Risk)#230
length(CaseControl_c) #230
bach<-cbind(Risk, CaseControl_c)
bach<-data.frame(bach)
attach(bach)
```

```
######
par(mfrow = c(1,2))
library(pROC)
library(ROCR)
library(gdata)
library(Daim)
library(verification)
#Plots a ROC for a given model
roc.plot(bach$CaseControl_c,bach$Risk, xlab = "False positive rate",
ylab = "True positive rate", main = NULL, CI = T, n.boot = 100,
plot = "both", binormal = TRUE)
M <- roc(bach$Risk,bach$CaseControl_c, "1")
summary(M)
plot(M,main = NULL)
#Sensitivity : 0.693 ; Specificity :  0.490 ; best cut-off :  0.010 ;  AUC :  0.606
ci.auc(bach$CaseControl_c,bach$Risk) # 95% CI: 0.517-0.695 (DeLong)
##### AUC #####
auc = colAUC(bach$Risk,bach$CaseControl_c, plotROC=TRUE, alg=c("Wilcoxon","ROC"))
auc  #0 vs. 1   0.6059
# Confidence Interval
t <- wilcox.test(bach$Risk,bach$CaseControl_c,paired=FALSE,conf.int=0.95)
######## Calcuating Sensitivity- Specificity ########
library(caret) #to calculate Sensitivity / Positivity
for (i in 1:230)
{if (bach$Risk[i]<0.025) {bach$Risk[i]<-0}
{if (bach$Risk[i]>=0.025) {bach$Risk[i]<-1}}}  #Recode to Case=1 and Control=0
sensitivity(factor(bach$Risk), factor(bach$CaseControl_c)) #0.8824
specificity(factor(bach$Risk), factor(bach$CaseControl_c)) #0.1564
posPredValue(factor(bach$Risk), factor(bach$CaseControl_c)) # 0.2296
negPredValue(factor(bach$Risk), factor(bach$CaseControl_c)) #0.8235
#### LLP Model ####
# Assign "Nil" to unknown values
for(i in 1:length(AgeRegistered))
{if (is.na(AgeRegistered[i])) {AgeRegistered[i]=0}
{if (is.na(Sex[i])) {Sex[i]="NA"} }}
```



```
#AgeRegistered = 5x+y
y<-rep(NA,length(AgeRegistered))
for(i in 1:length(AgeRegistered))
{y[i]<-AgeRegistered[i]%%5} # Age Registered Modulus 5
x<-rep(NA,length(AgeRegistered))
for(i in 1:length(AgeRegistered))
{x[i]<-(AgeRegistered[i]-y[i])/5} # Age Registered multiple of 5
AgeRegistered[1] #46.4
x[1] #9
y[1] #1.4
### Calculating Alpha ###
alpha<-rep(NA,879)
for(i in 1:length(AgeRegistered))
{if (AgeRegistered[i]>=40 && AgeRegistered[i]<45 && Sex[i]=="2") {alpha[i]<--9.9}
{if (AgeRegistered[i]>=45 && AgeRegistered[i]<50 && Sex[i]=="2") {alpha[i]<--8.06}
{if (AgeRegistered[i]>=50 && AgeRegistered[i]<55 && Sex[i]=="2") {alpha[i]<--7.46}
{if (AgeRegistered[i]>=55 && AgeRegistered[i]<60 && Sex[i]=="2") {alpha[i]<--6.5}
{if (AgeRegistered[i]>=60 && AgeRegistered[i]<65 && Sex[i]=="2") {alpha[i]<--6.22}
{if (AgeRegistered[i]>=65 && AgeRegistered[i]<70 && Sex[i]=="2") {alpha[i]<--5.99}
{if (AgeRegistered[i]>=70 && AgeRegistered[i]<75 && Sex[i]=="2") {alpha[i]<--5.49}
{if (AgeRegistered[i]>=75 && AgeRegistered[i]<80 && Sex[i]=="2") {alpha[i]<--5.23}
{if (AgeRegistered[i]>=80 && AgeRegistered[i]<85 && Sex[i]=="2") {alpha[i]<--5.42}
{if (AgeRegistered[i]>=40 && AgeRegistered[i]<45 && Sex[i]=="1") {alpha[i]<--9.06}
{if (AgeRegistered[i]>=45 && AgeRegistered[i]<50 && Sex[i]=="1") {alpha[i]<--8.16}
{if (AgeRegistered[i]>=50 && AgeRegistered[i]<55 && Sex[i]=="1") {alpha[i]<--7.31}
{if (AgeRegistered[i]>=55 && AgeRegistered[i]<60 && Sex[i]=="1") {alpha[i]<--6.63}
{if (AgeRegistered[i]>=60 && AgeRegistered[i]<65 && Sex[i]=="1") {alpha[i]<--5.97}
{if (AgeRegistered[i]>=65 && AgeRegistered[i]<70 && Sex[i]=="1") {alpha[i]<--5.56}
{if (AgeRegistered[i]>=70 && AgeRegistered[i]<75 && Sex[i]=="1") {alpha[i]<--5.31}
{if (AgeRegistered[i]>=75 && AgeRegistered[i]<80 && Sex[i]=="1") {alpha[i]<--4.83}
{if (AgeRegistered[i]>=80 && AgeRegistered[i]<85 && Sex[i]=="1") {alpha[i]<--4.68}
}}}}}}}}}}}}}}}}}}}}
for(i in 1:length(AgeRegistered))
{alpha[i]<-( (5-y[i]-0.5)*alpha[i] + (y[i]+0.5)*alpha[i+5] )/5}
### Calculating Family history for Beta ###  ### Beta ###
```

```
beta1<-rep(NA,879)
for(i in 1:879)
{if (dur[i]==0) {beta1[i]<-0}
{if (dur[i]>=1 && dur[i]<20) {beta1[i]<-0.769}
{if (dur[i]>=21 && dur[i]<40) {beta1[i]<-1.452}
{if (dur[i]>=41 && dur[i]<60) {beta1[i]<-2.507}
{if (dur[i]>=60) {beta1[i]<--2.724}}}}}
beta2<-rep(NA,879)
for(i in 1:879)
{if (S20[i]=="Yes") {beta2[i]<-0.602}
else {beta2[i]<-0} }
### Calculate Early and Late inset familty values
for(i in 1:879) #Assign "0" to NA values
{if (is.na(Family[i])) {Family[i]<-0}}
for(i in 1:879)
{if (AgeRegistered[i]>60 &&  Family[i]=="1") {Family[i]=1}  # Late onset then Family=1
{if (AgeRegistered[i]>60 &&  Family[i]=="2") {Family[i]=1}  # Late onset then Family=1
{if (AgeRegistered[i]<60 &&  Family[i]=="1") {Family[i]=2}  # Early onset then Family=2
{if (AgeRegistered[i]<60 &&  Family[i]=="2") {Family[i]=2}  # Early onset then Family=2}}}}
beta3<-rep(NA,879)
for(i in 1:879)
{if (is.na(Family[i])) {beta3[i]<-0}
{if (Family[i]=="2") {beta3[i]<-0.703} #Early
{if (Family[i]=="1") {beta3[i]<-0.168} #Late}}}
beta<-rep(NA,879) #Create an array of NAs
for (i in 1:879)
{beta[i]<-beta1[i]+beta2[i]+beta3[i]}
### Calculating P ###
p<-rep(NA,879) #Create an array of NAs
for(i in 1:879)
{p[i]= (1/ (1+exp(- ( alpha[i] +beta[i] ))))}
summary(p) # mean =1.802 e-02
#Exclude unknown values of p and respectively uknown values of Case Control
for(i in 1:879)
{if (is.na(p[i])) {CaseControl_c[i]=NA}}
```



```
p<-(na.exclude(p))
CaseControl_c<-(na.exclude(CaseControl_c))
#for(i in 1:length(alpha))
#{if (is.na(alpha[i])) {alpha[i]=0}}
CaseControl_c <-ifelse(CaseControl_c=="Case", 1, 0)
#Recode to Case=1 and Control=0
length(p) #106
length(CaseControl_c)  #106
llp<-cbind(p,CaseControl_c)
llp<-data.frame(llp)
attach(llp)
par(mfrow = c(1,2))
library(ROCR)
library(gdata)
library(Daim)
library(verification)
#Plots a ROC for a given model
roc.plot(llp$CaseControl_c, llp$p, xlab = "False positive rate",
ylab = "True positive rate", main = NULL, CI = T, n.boot = 100,
 plot = "both", binormal = TRUE)
M <- roc(llp$p, llp$CaseControl_c, "1")
summary(M)
plot(M) #Sensitivity : 0.619 ; Specificity : 0.864 ; best cut-off : 0.01043
auc = colAUC(llp$p,llp$CaseControl_c, plotROC=TRUE, alg=c("Wilcoxon","ROC"))
auc  #0 vs. 1  0.7192
#http://rss.acs.unt.edu/Rdoc/library/verification/html/roc.plot.html
#CIs are computed with Delong's method (DeLong et al.) based
on U-statistics theory and asymptotic normality. As this test does not
# require bootstrapping
ci.auc(llp$CaseControl_c,llp$p) #
######## Calcuating Sensitivity- Specificity ########
library(caret) #to calculate Sensitivity / Positivity
for (i in 1:106)
{if (llp$p[i]<0.0104) {llp$p[i]<-0}
{if (llp$p[i]>=0.0104) {llp$p[i]<-1}}}  #Recode to Case=1 and Control=0
```

```
sensitivity(factor(llp$p), factor(llp$CaseControl_c) ,"1") # 0.619
posPredValue(factor(llp$p), factor(llp$CaseControl_c)) #0.3725
negPredValue(factor(llp$p), factor(llp$CaseControl_c)) #0.9455
```

11.3 Research Questionnaire

Scale document down

Cancer Research Centre
Weston Park Hospital (STH)
The University of Sheffield

ReSoLuCENT

Research Questionnaire

This questionnaire provides us with important information for our study. We need to know some details about your own lifestyle and exposure to risk factors. Then we will ask similar questions about your relatives.

All the information which we collect about you during the course of the research will be kept strictly confidential.

Full name _____ _____
Forename(s) Last name

Any other names you have been known by: _____

Your NHS No - *This is a 10 digit number, you may find it on letters from the hospital or an appointment card*

Address

GP Name _____

GP Address _____

Your Post code

GP Postcode

Your Date of Birth

If you are unsure about any of the questions you will have the opportunity to discuss it with the nurse when you have your blood sample taken.

Thank you for completing this questionnaire.

For office use only

Study reference number

Affix barcode label

Figure 26:

About you

We will begin the interview by asking you some questions about yourself. This will include questions about where you have lived, where you have worked, and questions about your health.

S1. Where were you born? Town _____

Country _____

S2. How long have you lived at your current address? ☐☐ years

S3. How long have you lived in this area? (within a 25 mile radius) ☐☐ years

S4. How many people live at home with you? (not including yourself). ☐☐

Who are they? Please give numbers not names

Others, please specify eg lodger, flatmate

☐ Partner/spouse ☐ Parents ☐ _____

☐ Children/stepchildren ☐ Brothers/ sisters ☐ _____

About your health

S5. How would you describe your own health at the moment?

☐ Excellent ☐ Good ☐ OK ☐ Poor ☐ Very poor

S6. What is your height? ☐ ft ☐☐ ins **or** ☐☐☐ cms

S7. Approximately how much did you weigh when you were 20 yrs old?

☐☐ Stones ☐☐ lbs **or** ☐☐☐·☐ kg

S8. Approximately how much did you weigh 1 year ago?

☐☐ Stones ☐☐ lbs **or** ☐☐☐·☐ kg

Figure 27:

About your health

S9. Have you ever suffered from any of the following?

Lung Diseases			**Age at onset**
Asthma	☐ Yes	☐ No	☐☐
Bronchitis	☐ Yes	☐ No	☐☐
Emphysema	☐ Yes	☐ No	☐☐
Pneumonia	☐ Yes	☐ No	☐☐
Coal workers' pneumoconiosis	☐ Yes	☐ No	☐☐
Tuberculosis	☐ Yes	☐ No	☐☐
Other _____	☐ Yes	☐ No	☐☐
Heart and circulation diseases			
High blood pressure (requiring drug treatment)	☐ Yes	☐ No	☐☐
Angina	☐ Yes	☐ No	☐☐
Heart attack	☐ Yes	☐ No	☐☐
Stroke	☐ Yes	☐ No	☐☐
Other _____	☐ Yes	☐ No	☐☐

Cancer & other malignant diseases (includes leukaemia, lymphoma etc) ☐ Yes ☐ No

Have you had more than one cancer? ☐ Yes ☐ No

For each cancer please state:

What part of the body was affected?
Your age at diagnosis
What type of treatment did you have? eg surgery, radiotherapy, chemotherapy

_____

_____

_____

_____

_____

Any other long term health problem? ☐ Yes ☐ No Age at diagnosis/onset? ☐☐

What is the problem? _____

What type of treatment? _____

Figure 28:

About your work and education

We are going to ask you some questions about your education and employment

S10. How old were you when you finished full-time school or college? ☐☐ yrs old

S11. What educational qualifications do you have - tick all that apply

☐ None	☐ City & Guilds	☐ Secretarial college
☐ School leaving certificate	☐ NVQ 1	☐ Teaching diploma
☐ CSE	☐ NVQ 2	☐ University degree
☐ O level	☐ NVQ 3	☐ Post-graduate degree
☐ GCSE	☐ NVQ 4	☐ Other, please give details
☐ Matriculation	☐ HND	_____
☐ A level or Highers	☐ Completed apprenticeship	
☐ Technical college exams	☐ Trade certificate	_____

S12. At present are you:

☐ Employed ☐ Chronic sick ☐ Unemployed due to disability ☐ Student
☐ Retired ☐ Unemployed ☐ Home duties

The following questions refer to your current main job, or (if you are not working now) to your last main job.
Please tick one box only per question.

S13. Do (did) you work as an employee or are (were) you self-employed?

Employee ☐
Self-employed with employees ☐
Self-employed / freelance without employees (go to question 16) ☐

S14. About the size of the organisation

For employees: indicate below how many people work (worked) for your employer at the place where you work (worked).

For self-employed: indicate below how many people you employ (employed). Go to question 16 when you have completed this question.

1 to 24 ☐
25 or more ☐

S15. Do (did) you supervise any other employees?

Eg. A supervisor or foreman is responsible for overseeing the work of other employees on a day-to-day basis

☐ Yes ☐ No

Figure 29:

About your work

S16. Your occupation

Please tick one box to show which best describes the sort of work you do. (If you are not working now, please tick a box to show what you did in your last job).

PLEASE TICK ONE BOX ONLY

Modern professional occupations ☐
such as: teacher - nurse - physiotherapist - social worker - welfare officer - artist - musician - police officer - software designer

Clerical and intermediate occupations ☐
such as: secretary - personal assistant - clerical worker - office clerk - call centre agent - nursing auxiliary - nursery nurse

Senior managers or administrators ☐
(usually responsible for planning, organising and co-ordinating work and for finance) such as: finance manager - chief executive

Technical and craft occupations ☐
such as: motor mechanic - fitter - inspector - plumber - printer - tool maker - electrician - gardener - train driver

Semi-routine manual and service occupations ☐
such as: postal worker - machine operative - security guard - caretaker - farm worker - catering assistant - receptionist - sales assistant

Routine manual and service occupations ☐
such as: HGV driver - van driver - cleaner - porter - packer - sewing machinist - messenger - labourer - waiter / waitress - bar staff

Middle or junior managers ☐
such as: office manager - retail manager - bank manager - restaurant manager - warehouse manager - publican

Traditional professional occupations ☐
such as: accountant - solicitor - medical practitioner - scientist - civil / mechanical engineer

About your ethnic group

S17. To which of these groups do you consider that you belong?

☐ White
☐ Black, Caribbean
☐ Black, Other Please give details _____
☐ Indian
☐ Pakistani
☐ Bangladeshi
☐ Chinese
☐ Other Please give details _____

Figure 30:

About your work

S18. Please list all your current and past occupations with start and end years only - you do not need to mention each employer or job separately if the type of work was the same

Job title	Employer activity	From	To
eg Miner	Coal industry	1963	1985
eg Bus driver	Public transport	1987	2001

S19. Have you ever worked in dusty conditions? (If yes please give details) ☐Yes ☐No

S20. Have you ever worked with asbestos? (If yes please give details) ☐Yes ☐No

S21. Have you ever been exposed to fumes at work? (If yes please give details) ☐Yes ☐No

S22. Have you ever been exposed to toxic chemicals at work or through an industrial accident? (If yes please give details) ☐Yes ☐No

S23. Is there anything that you think may be relevant eg exposure to radon, living near an incinerator or chemical factory (If yes please give details) ☐Yes ☐No

Figure 31:

About your lifestyle

Alcohol

S24. Have you ever drunk alcohol?

☐No, I've always been teetotal - Go to question S28
☐Only on special occasions - Go to question S28
☐More than occasionally - please answer question S25

S25. When you were aged 20, about how many alcoholic drinks did you have each week?

Put "0" if none or less than one drink a week
Please answer EACH line

Beer or cider	☐☐☐	pints each week
Wine	☐☐☐	glasses each week
Sherry or other fortified wine	☐☐☐	glasses each week
Spirits	☐☐☐	glass (singles) each week

S26. When you were aged 40, about how many alcoholic drinks did you have each week?

Put "0" if none or less than one drink a week
Please answer EACH line

Beer or cider	☐☐☐	pints each week
Wine	☐☐☐	glasses each week
Sherry or other fortified wine	☐☐☐	glasses each week
Spirits	☐☐☐	glass (singles) each week

S27. One year ago, about how many alcoholic drinks did you have each week?

Put "0" if none or less than one drink a week
Please answer EACH line

Beer or cider	☐☐☐	pints each week
Wine	☐☐☐	glasses each week
Sherry or other fortified wine	☐☐☐	glasses each week
Spirits	☐☐☐	glass (singles) each week

Figure 32:

Smoking

S28. Have you ever smoked *regularly*? ie as much as 10 cigarettes per week for as long as a year ☐ Yes ☐ No

If 'No' please go to question S33

If 'Yes' are you currently ☐ An ex-smoker ☐ A smoker

S29. How old were you when you had your first full cigarette? ☐☐ yrs old

S30. How old were you when you started smoking cigarettes *regularly*? ☐☐ yrs old

S31. **If you have stopped smoking**, how old were you when you last smoked? ☐☐ yrs old

Did you smoke at the following ages? If so, how many cigarettes did you smoke and were they usually filter cigarettes?

Age 20 ☐☐☐ cigs per day ☐ Filter ☐ No filter ☐ Non-smoker

Age 30 ☐☐☐ cigs per day ☐ Filter ☐ No filter ☐ Non-smoker

Age 40 ☐☐☐ cigs per day ☐ Filter ☐ No filter ☐ Non-smoker

Age 50 ☐☐☐ cigs per day ☐ Filter ☐ No filter ☐ Non-smoker

1 yr ago ☐☐☐ cigs per day ☐ Filter ☐ No filter ☐ Non-smoker

S32. How deeply did you inhale?

☐ Deeply into the lungs ☐ A little ☐ Not at all

What kind of cigarettes did you usually smoke? (eg filter, low tar, roll your own, what brand)

_____

_____

Figure 33:

Other tobacco use

S33. Did you smoke cigars one year ago? ☐ Yes ☐ No

If Yes, how many cigars did you smoke per week? ☐☐ **Or** per day? ☐☐

How long have you smoked cigars? ☐☐ yrs

S34. Did you smoke a pipe one year ago? ☐ Yes ☐ No

If Yes, how much pipe tobacco did you consume per week? ☐☐ **oz** ☐☐☐ **g**

How long have you smoked a pipe? ☐☐ yrs

S35. Have you ever used any other tobacco-related products (such as chewing tobacco, cannabis, snuff etc)

☐ No - Go to question S36

☐ Yes - please give details below

Product	Years of use	Frequency/amount

Figure 34:

Other people's smoking

S36. One year ago how many of the people living in your home smoked? (not including yourself) ☐☐

Who are they? Please give numbers not names

		Others, please specify eg lodger, flatmate
☐ Partner/spouse	☐ Parents	☐ _____
☐ Children/stepchildren	☐ Brothers/ sisters	☐ _____

S37. Are you currently exposed to other people's smoke through your work or leisure activities?

☐ No - Go to question S38

☐ Yes - please give details

Approximate number of hours per week ☐☐☐ hrs

Details _____

_____

S38. When you were a child how many of the people who lived in your home smoked? (not including yourself) ☐☐

Who were they? Please give numbers not names

		Others, please specify eg lodger, flatmate
☐ Parents/step-parents	☐ Brothers/sisters	☐ _____
☐ Grandparents	☐ Aunts/uncles	☐ _____

Figure 35:

About your parents

F1. Is your father still alive? ☐ Yes ☐ No

F1a. If your father is dead what was the cause of his death?

_____

F2. What age is your father now, or what age was he when he died. ☐☐ yrs old

F3. Is/was your father ☐ A non-smoker ☐ An ex-smoker ☐ A smoker

F4. Has your father ever suffered from any of the following?

				Age at onset
Lung Diseases				
Asthma	☐ Yes	☐ No	☐ Unknown	☐☐
Bronchitis	☐ Yes	☐ No	☐ Unknown	☐☐
Emphysema	☐ Yes	☐ No	☐ Unknown	☐☐
Pneumonia	☐ Yes	☐ No	☐ Unknown	☐☐
Coal workers' pneumoconiosis	☐ Yes	☐ No	☐ Unknown	☐☐
Tuberculosis	☐ Yes	☐ No	☐ Unknown	☐☐
Other _____	☐ Yes	☐ No	☐ Unknown	☐☐
Heart and circulation diseases				
High blood pressure (requiring drug treatment)	☐ Yes	☐ No	☐ Unknown	☐☐
Angina	☐ Yes	☐ No	☐ Unknown	☐☐
Heart attack	☐ Yes	☐ No	☐ Unknown	☐☐
Stroke	☐ Yes	☐ No	☐ Unknown	☐☐
Other _____	☐ Yes	☐ No	☐ Unknown	☐☐
Cancer & other malignant diseases (includes leukaemia, lymphoma etc)	☐ Yes	☐ No	☐ Unknown	
Has he had more than one cancer?	☐ Yes	☐ No	☐ Unknown	

For each cancer please state:
- What part of the body was affected?
- His age at diagnosis
- What type of treatment did he have? eg surgery, radiotherapy, chemotherapy

_____

_____

_____

Any other long term health problem? ☐ Yes ☐ No ☐ Unknown

Age at diagnosis/onset? ☐☐

What is/was the problem? _____

What type of treatment? _____

Figure 36:

About your parents

M1. Is your mother still alive? ☐ Yes ☐ No

M1a. If your mother is dead what was the cause of her death?

_____

M2. What age is your mother now, or what age was she when she died? ☐☐ yrs old

M3. Is/was your mother ☐ A non-smoker ☐ An ex-smoker ☐ A smoker

M4. Has your mother ever suffered from any of the following?

				Age at onset
Lung Diseases				
Asthma	☐ Yes	☐ No	☐ Unknown	☐☐
Bronchitis	☐ Yes	☐ No	☐ Unknown	☐☐
Emphysema	☐ Yes	☐ No	☐ Unknown	☐☐
Pneumonia	☐ Yes	☐ No	☐ Unknown	☐☐
Coal workers' pneumoconiosis	☐ Yes	☐ No	☐ Unknown	☐☐
Tuberculosis	☐ Yes	☐ No	☐ Unknown	☐☐
Other _____	☐ Yes	☐ No	☐ Unknown	☐☐
Heart and circulation diseases				
High blood pressure (requiring drug treatment)	☐ Yes	☐ No	☐ Unknown	☐☐
Angina	☐ Yes	☐ No	☐ Unknown	☐☐
Heart attack	☐ Yes	☐ No	☐ Unknown	☐☐
Stroke	☐ Yes	☐ No	☐ Unknown	☐☐
Other _____	☐ Yes	☐ No	☐ Unknown	☐☐
Cancer & other malignant diseases (includes leukaemia, lymphoma etc)	☐ Yes	☐ No	☐ Unknown	
Has she had more than one cancer?	☐ Yes	☐ No	☐ Unknown	

For each cancer please state:
- What part of the body was affected?
- Her age at diagnosis
- What type of treatment did she have? eg surgery, radiotherapy, chemotherapy

_____

_____

_____

Any other long term health problem? ☐ Yes ☐ No ☐ Unknown

Age at diagnosis/onset? ☐☐

What is/was the problem? _____

What type of treatment? _____

Figure 37:

Information about your brothers

B1. How many brothers do/did you have? ☐☐

B2. For each brother complete the following

Year of birth	Relative full/ half/ step/ adopted	Are they still alive?	Age now or at death	Smoking status Non/ Ex/ Smoker	Have they ever been diagnosed with cancer?	If yes, what part of the body was affected?	Age at onset/ diagnosis	Type of treatment etc

Figure 38:

Information about your sisters

Si1. How many sisters do/did you have? ☐☐

Si2. For each sister complete the following

Year of birth	Relative full/ half/ step/ adopted	Are they still alive?	Age now or at death	Smoking status Non/ Ex/ Smoker	Have they ever been diagnosed with cancer?	If yes, what part of the body was affected?	Age at onset/ diagnosis	Type of treatment etc

Figure 39:

Figure 40:

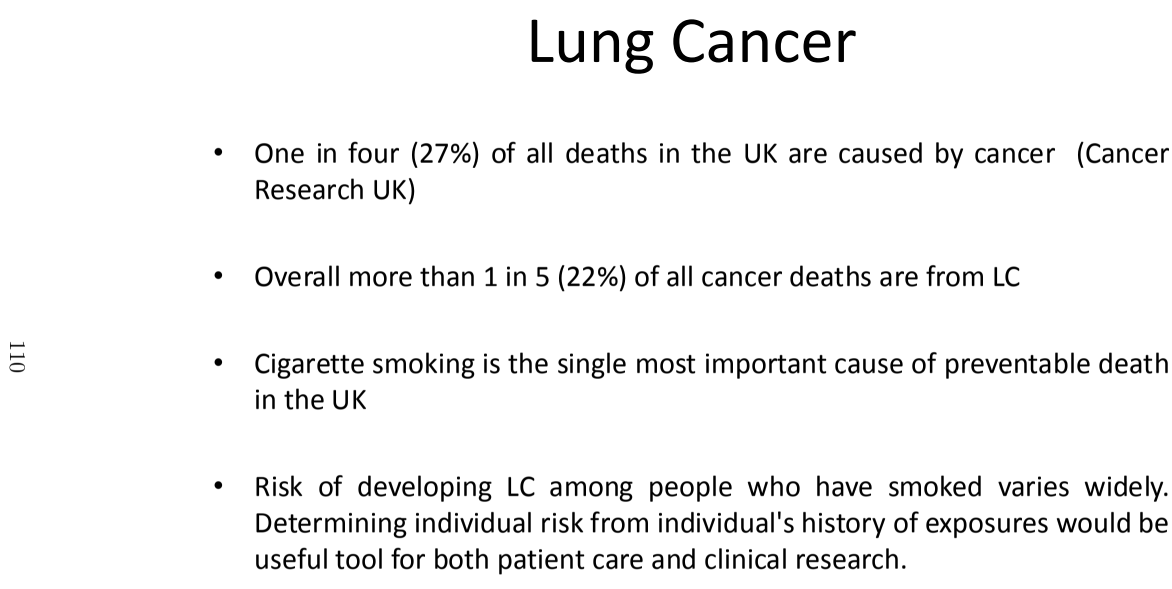

Figure 41:

Resolucent Data

- Resource for the Study of Lung Cancer Epidemiology in North Trent (ReSoLuCENT) is a study in the NIHR CRN (National Institute for Health Research Clinical Research Network)
- Trial Location: Airedale ,Cardiff, Manchester, Sheffield, Southampton
- Eligible patients are identified through Lung Multidisciplinary Teams or clinic and the Consultant agrees
- Patient complete the Patient Information Sheet
- Patients and controls complete the Lifestyle Questionnaire

Figure 42:

Eligibility Criteria

- Total 879 subjects : 516 patients and 363 controls recruited from April 2006 to March 2011
- Cases have LC or need to have an operation for suspected LC
  - Are 60 years old or less **or**
  - Have a first degree relative aged 60 or less who has LC **or**
  - Have two or more first or second degree relatives of any age with LC
- Controls were cancer free
  - Recruited through Cases, are the partner of someone taking part in the study **or**
  - Are a first degree relative of someone taking part in the study, and are at least 18 years old

Figure 43:

Exploratory Data Analysis

- Crude data: 24 variables used such as: Case Control, Age Registered , Sex, Death Date, Date Diagnosis, Site of LC
- Processed data: Pack Year, Smoking Duration, Smoking Cessation Duration (ex-smokers), Cigarettes Per Day (CPD)
- Mean age for cases and controls are 55.27 and 51.27 years respectively
- A regular smoker is defined as smoking as much as 10 cigarettes per week for a year
- 249 Current Smoker, 345 Ex-smoker and 154 Non-smoker

Figure 44:

Pack Years

- Pack Years is accumulated summary measure of cigarette smoking over someone's lifetime, calculated as the number of packs per day times the number of years a person has smoked.
- Number of pack years = (number of cigarettes smoked per day x number of years smoked)/20
    - 20 cigarettes per pack

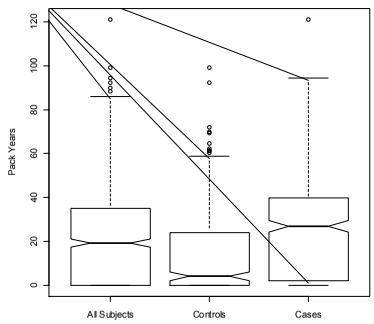

Figure 45:

Risk Models

- Risk Assessment Models estimate binary outcome of developing cancer over a defined period of time
- Cancer risk prediction models help identify high risk individuals and control suffering and death by preventive methods
- Models also facilitate the design and planning of clinical chemoprevention trials, development of benefit-risk indices and provide estimates of the population burden and cost of cancer
- Model inputs : Subject's gender, sex, family history, asbestos exposure history and smoking history (Cigarettes per day, smoking duration, smoking cessation duration)

Figure 46:

Lung Cancer Prediction

- Two risk prediction models have been proposed.
  - Bach : 10 year LC prediction model from a large, multicentre, randomized, controlled trial of LC using S-Plus, published in 2003
  - LLP : 5 year absolute risk model for case-control study in liver pool area using STATA, 2008
  - D'amelio : Compared Bach and LLP discriminatory power using 5-year LC risk using MATLAB, 2010
- The 2 models share risk factors and vary in the inclusion of lung related co morbidities or family history.
- Objective : Apply models to dataset to assess prediction accuracy

Figure 47:

Bach Model

- 18172 subjects enrolled in the Carotene and Retinol Efficacy Trial (CARET) a multicenter, randomized, controlled study from 1985 to 1994
  - Continuous predictors : Age, duration of abstinence, duration of smoking, and number of cigarettes smoked per day
  - Categorical variables : Sex, and asbestos exposure
- Is predictive only for specific age band, cigarettes smoked and certain number of years
- Regression analysis yields multivariate 1-year models :
  - Probability of being diagnosed with LC (focus of study)
  - Probability that an individual will die without having been diagnosed with LC (competing risk)
- Recursively estimated 10-year LC risk by cycling these two 1-year models 10 times
- Result : The risk of LC varies widely among smokers

Figure 48:

Liverpool Lung Project Model

- 1736 cases and controls recruited between 1998 and 2005
- Accounts : Age, sex and smoking (self reported pack year) , family history of LC, occupational exposure to asbestos, prior diagnosis of pneumonia and prior diagnosis of a malignant tumour other than LC
- Conditional logistic regression, backward stepwise regression performed on multivariate model, then converted to absolute risk model
- LLP risk model discriminates between high and low risk individuals

- Result : LLP is simple and more directly applicable for use in the primary care setting. Could predict approximately two-thirds of LC within 5-years, screening only 30% of the population

Figure 49:

D’amelio Study

- 4900 subjects recruited for case–control study at the Harvard School of Public Health and Massachusetts General Hospital
- Estimated 5-year LC risk for Bach and LLP risk model and compared the discriminatory power, accuracy, and clinical utility of these models
- Models’ discriminatory power compared by calculating the area under the curve (AUC) of the receiver operator characteristic (ROC) curve
  - LLP had comparable discriminatory power (0.69), whereas the Bach model had significantly lower power (0.66)
- Negative predictive values (probability of accurately categorising an unaffected participant) were higher with the LLP model
- Positive predictive value (probability of categorising an affected person who actually have disease) were higher with Bach model
- Bach models had lower sensitivity but better specificity than LLP model

119

Figure 50:

Conclusion

Overall, the Liverpool Lung Project (LLP) model had a high discriminatory power 0.72, whereas the Bach model had significantly lower power 0.60. Positive predictive values were slightly higher with Bach model but negative predictive values were higher for the LLP models. The LLP model had lower sensitivity but better specificity than did the Bach model.

120

Moderate differences were observed in discriminatory power among the lung cancer risk models confirming the difficulty in developing effective risk models. LLP model was favourable over Bach model for higher discriminatory power and inclusion of family history.

Figure 51: