

RICE UNIVERSITY

**Lung Carcinogenesis Modeling: Resampling and Simulation
Approach to Model Fitting, Validation, and Prediction**

by

Millennia Foy

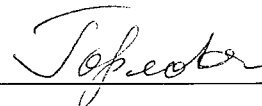
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

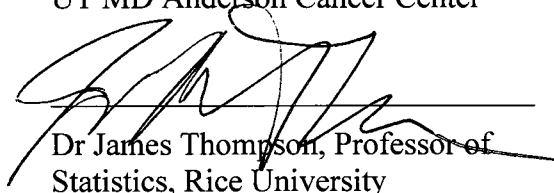
APPROVED THESIS COMMITTEE:



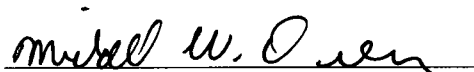
Dr Marek Kimmel, Chair
Thesis Co-director
Professor of Statistics, Rice University



Dr Olga Gorlova, Thesis Director
Assistant Professor of Epidemiology
UT MD Anderson Cancer Center



Dr James Thompson, Professor of
Statistics, Rice University



Dr Michael Deem, John W. Cox
Professor of Biochemical and Genetic
Engineering, Rice University

HOUSTON, TX
NOVEMBER 2009

UMI Number: 3421308

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

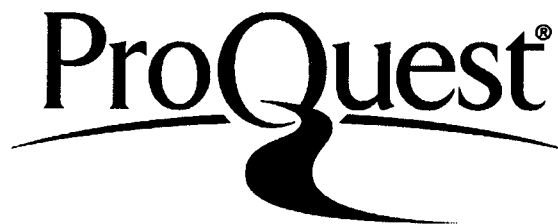
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3421308

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

Lung Carcinogenesis Modeling: Resampling and Simulation Approach to Model Fitting,
Validation, and Prediction

by

Millennia Foy

Because of serious health implications, lung cancer is the leading cancer killer for both men and women. It is well known that smoking is the major risk factor for lung cancer. I propose to use a two-stage clonal expansion (TSCE) model to evaluate the effects of smoking on initiation and promotion of lung carcinogenesis.

The TSCE model is traditionally fit to prospective cohort data. A new method has been developed that allows reconstruction of cohort data from the combination of risk factor data from a case-control study, and tabulated incidence/mortality rate data. A simulation study of the method shows that it is accurate in estimating the parameters of the TSCE model.

The method is then applied to fit a TSCE model based on smoking history. The fitted model is then validated in two ways. First the model is used to predict lung cancer deaths in the non-asbestos exposed control arm of the CARET study, where the model predicts 366.8 lung cancer deaths while there were 364 observed. Second, the model is used to simulate LC mortality in the US population and reasonably reproduced observed US mortality rates.

The model is also applied to a study of CT screening for lung cancer. The study is a single arm CT screening study lacking a control arm for comparison. The model is used to simulate LC mortality in the absence of screening to serve as a surrogate control arm for comparison. Based on the model there is a statistically significant mortality reduction of 36% due to CT screening.

Acknowledgements

This research was supported, in part, by a cancer prevention fellowship from the National Cancer Institute training grant R25T CA57730, Robert M. Chamberlain, Ph.D., Principal Investigator, University of Texas M.D. Anderson Cancer Center, and also by a T32 training grant from the National Cancer Institute training grant T32 5T32CA096520, Gary L. Rosner, Ph.D., Principal Investigator, University of Texas M.D. Anderson Cancer Center.

Table of Contents

ABSTRACT.....	ii
Acknowledgements.....	iv
Table of Contents.....	v
List of Figures.....	i
List of Tables.....	i
Chapter 1 Background and Significance.....	1
1.1 Introduction.....	1
1.2 Epidemiology of Lung Cancer.....	3
1.3 Temporal Trends in Lung Cancer Mortality.....	4
1.4 Risk Factors for Lung Cancer.....	6
1.4.1 <i>Smoking Tobacco</i>	7
1.4.2 <i>Secondhand Smoke</i>	7
1.4.3 <i>Environmental Exposures</i>	8
1.4.4 <i>Pre-existing Lung Disease</i>	9
1.4.5 <i>Family History of Cancer</i>	9
1.4.6 <i>Genetic Susceptibility of lung cancer, DNA repair capacity</i>	10
1.5 Attributable Risks.....	11
1.6 Gender Differences in Lung Cancer.....	11
1.7 Lung Cancer in Never Smokers.....	13
1.8 Screening for Lung Cancer.....	15
1.8.1 <i>CT Screening for LC</i>	15
Chapter 2 Carcinogenesis Modeling.....	18
2.1 Review of Carcinogenesis Models.....	18
2.1.1 <i>Two-stage Carcinogenesis Models</i>	19
2.2 Two-stage clonal expansion (TSCE) Model.....	20
2.2.1 <i>Non-identifiability in the TSCE Model</i>	21
2.2.2 <i>Evolution of the TSCE model</i>	22
2.2.3 <i>TSCE Models Applied to Lung Cancer</i>	23
2.3 Fitting the TSCE Model.....	24
2.3.1 <i>Prospective Cohort-based Likelihood</i>	24
2.3.2 <i>Re-sampling Based Method of Reconstructing Cohort Data</i>	24
2.3.3 <i>Re-sampling Method Applied to a Simple Model</i>	26
Chapter 3 Simulation Studies.....	28
3.1 Simulation Methods.....	28
3.2 Simulated TSCE Parameters.....	30
3.3 Results for Traditional Prospective Cohort Method.....	31
3.4 Results for Resampling Method.....	33
Chapter 4 Smoking Based TSCE Model and Validation.....	36
4.1 MD Anderson Case-Control Data.....	36
4.2 Sources of Age-specific Mortality.....	39
4.2.1 <i>Cancer Prevention Study 1</i>	39
4.2.2 <i>Nurses Health Study and Health Professionals Follow-up Study</i>	40
4.2.3 <i>Comparisons of Mortality Datasets</i>	41

4.2.4	<i>National Health Interview Survey</i>	44
4.3	TSCE Model based on Smoking History	44
4.4	Predicting LC Mortality in the CARET study	48
4.5	Simulating US LC mortality	56
4.6	CISNET Smoking Base Case Project	59
Chapter 5	Implications for Screening	63
5.1	Single-arm CT Screening Trial for LC	63
5.2	NY LC Screening Study	63
5.3	Simulating Mortality in the Absence of Screening.....	64
5.4	Results of Screening Simulation.....	66
5.5	CT Screening Controversy.....	68
Chapter 6	Discussion and Future Directions	70
6.1	Summary	70
6.2	Future Considerations	71
References	74

List of Figures

Figure 1.1 Age-adjusted lung cancer mortality rates for males	5
Figure 1.2 Age-adjusted lung cancer mortality rates for females.....	6
Figure 2.1 Depiction of the TSCE model	20
Figure 2.2 Cohort likelihood applied to reconstructed pseudo-cohort data.....	27
Figure 3.1 Simulated age-specific incidence rates.....	31
Figure 3.2 Predicted incidence rates- traditional cohort method with biological parameters	32
Figure 3.3 Predicted incidence rates - traditional cohort method with identifiable parameters	32
Figure 3.4 Predicted incidence rates- resampling method with biological parameters	34
Figure 3.5 Predicted incidence rates- resampling method with identifiable parameters ..	34
Figure 4.1 Comparison of age-specific mortality rates for white males.....	42
Figure 4.2 Comparison of age-specific mortality rates for white females.....	42
Figure 4.3 Model predicted incidence rates	47
Figure 4.4 Yearly predicted and observed LC deaths- CARET	51
Figure 4.5 Cumulative predicted and observed LC deaths- CARET.....	51
Figure 4.6 Yearly predicted and observed LC deaths- CARET males	52
Figure 4.7 Cumulative predicted and observed LC deaths- CARET males	52
Figure 4.8 Yearly predicted and observed LC deaths- CARET females.....	53
Figure 4.9 Cumulative predicted and observed LC deaths- CARET females	53
Figure 4.10 Predicted and observed cumulative LC deaths- CARET (years 4-10).....	55
Figure 4.11 Predicted and observed cumulative LC deaths- CARET males (years 4-10)	55
Figure 4.12 Predicted and observed cumulative LC deaths- CARET females (years 4-10)	56
Figure 4.13 Simulated and observed US LC mortality rates- males.....	57
Figure 4.14 Simulated and observed US LC mortality rates- females	58
Figure 4.15 CISNET Smoking Base Case Simulation- males.....	60
Figure 4.16 CISNET Smoking Base Case Simulation- females.....	61
Figure 4.17 Proportion of LC deaths averted through tobacco control policies.....	62
Figure 5.1 Person-years in the NY cohort per year of follow-up	64
Figure 5.2 Yearly simulated and observed LC deaths in the NYC CT screening cohort .	66
Figure 5.3 Predicted and observed cumulative LC deaths in the NYC cohort (years 4-10)	67

List of Tables

Table 1.1 Median age at LC diagnosis by smoking status (Wakelee et al. 2007)	14
Table 2.1 Results of fitting the exponential model using the resampling method.....	27
Table 3.1 Results of fitting simulated data using traditional prospective cohort method.	31
Table 3.2 Results of fitting simulated data using the resampling method	33
Table 4.1 Characteristics of cases (n=3433) and controls (n=3132) available from the lung cancer case-control study (R01 CA55769, Spitz, PI)	37
Table 4.2 Pack-year histories for 992 white males and 919 white females included in analysis.....	38
Table 4.3 Results of family history analysis.....	38
Table 4.4 Cohort characteristics of mortality/incidence rate data	41
Table 4.5 Parameter fit for the TSCE model based on smoking	46
Table 4.6 CARET study predicted and observed LC deaths	50
Table 4.7 CARET predicted and observed LC deaths (years 4-10)	54
Table 5.1 Observed and predicted LC deaths in years 1-4 in the NYC cohort.....	66

Chapter 1

Background and Significance

1.1 Introduction

Lung cancer is the second leading cancer in terms of incidence for both men and women, second to prostate cancer for men and breast cancer for women. However, because of its serious health implications, lung cancer is the leading cancer killer for both men and women worldwide (Coleman et al 1993, NIH 2007). Once a patient is diagnosed, the prognosis is so poor that incidence data are often assumed to be equivalent to mortality data. Less than half of newly diagnosed cases live 1 more year. The 1-year survival rate increased from 37% in 1975 to 42% in 2000 while the 5-year survival rate for newly diagnosed cases is only 15% (Cancer Facts and Figures (CFF) 2009). The modest increase in 1-year survival could be due to advances in treatment such as surgical techniques, or the implementation of screening in some individuals. Although, screening for lung cancer is not yet recommended, some doctors are using chest x-ray or CT to screen for lung conditions in their patients. Lung cancer is a serious public health issue that needs to be studied in order to prevent lung cancer in those who do not have it, and improve survival in those who have been diagnosed

The goal of this project is to use carcinogenesis modeling, specifically the two-stage clonal expansion (TSCE) model in conjunction with maximum likelihood methods to estimate the effects of different risk factors on the development of lung cancer. Since the TSCE model is incidence based, it is normally fit to prospective cohort data. For this study, cohort data is unavailable but case-control data on risk factor exposure and tabled age-specific mortality rates are available. This prompted the development of a new method of fitting this model by reconstructing cohort data using re-sampling. A simulation study reveals that the proposed method is accurate in fitting the parameters of the TSCE model in the case where there are no known exposures.

Risk factors for lung cancer have been extensively researched. The main risk for lung cancer comes from tobacco smoke, including smoking but also to a lesser extent exposure to second-hand smoke. Other risk factors for lung cancer include exposure to radon, asbestos, and air pollution. Also, genetics plays a role as having a family history of lung cancer increases risk for lung cancer (NIH 2007). Findings from epidemiological studies on risk factors will be discussed in the background section.

In this thesis, I use a new resampling based method to reconstruct time to event data from the combination of case-control data and incidence/mortality rate data. Using the method, I fit a TSCE model based on smoking history. Using simulation, I validate the model against the heavy-smokers control arm of CARET. Also, using the CISNET smoking history generator, the model is able to simulate US LC mortality rates.

In 2006 as part of Li Deng's thesis and later published (Deng et al. 2009), the TSCE model was fit to MD Anderson case-control data on lung cancer using least squares estimation. Deng used the model to examine the effects of smoking and DRC on

lung carcinogenesis in current and never smokers exclusively. My model will differ not only in the method used for fitting, but on the risk factors included and the inclusion of former smokers.

Background information on lung cancer mortality trends, and known risk factors for lung cancer follow in the remainder of this chapter. The following chapter details the history of carcinogenesis models as well as describes the model for this study. It also describes the two-stage clonal expansion (TSCE) model and introduces a new method allowing for the model to be fit to case-control data instead of the traditional prospective cohort data. This thesis also includes simulation studies used to validate the accuracy of the method in fitting the parameters of the TSCE model. A fitted smoking based TSCE model is presented along with model validations. The model is then applied to a study of CT screening for lung cancer to determine the effectiveness in reducing lung cancer mortality. Discussion and future considerations are included in the final chapter.

1.2 Epidemiology of Lung Cancer

There are 2 major cytological types of lung cancer: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC). NSCLC makes up 87% of newly diagnosed lung cancers and is classified according to three different types. Squamous cell carcinoma makes up about 25% to 30% of all lung cancer diagnoses. This type of lung cancer tends to be located in the middle of the lungs near the bronchus. Adenocarcinoma accounts for 40% of lung cancer diagnoses and develops in the mucus producing glands of the lungs. It is the most common lung cancer among women and never smokers. Large-cell undifferentiated carcinoma tends to grow rapidly and can start in any part of

the lung. This type of lung cancer accounts for 10% to 15% of lung cancer diagnoses. SCLC makes up 13% of lung cancers and is characterized by small cells that multiply quickly forming large tumors that spread throughout the body quickly (NIH 2007).

1.3 Temporal Trends in Lung Cancer Mortality

As shown below the age-adjusted lung cancer mortality rates for men have been decreasing for almost 20 years while the women's rates have just stabilized. Women's rates of lung cancer are expected to also decline because their prevalence of smoking has been decreasing. African-American and white females have similar lung cancer mortality while African-American males have about a 50% increased mortality when compared to white males (Ries et al 1991). However, other races observed including: Hispanics, Native Americans, Alaska Natives and Pacific Islanders, have a significantly lower lung cancer mortality than whites and African-Americans. Mortality is chosen so that comparisons can be made between SEER and the Texas Cancer Information Center, which as of now only provides rates for mortality.

Surveillance Epidemiology and End Results (SEER) (National Cancer Institute) is a comprehensive database on cancer incidence in the United States. SEER provides data on cancer incidence, mortality and survival from 15 population-based cancer registries in the United States. SEER currently includes information on approximately 26% of the US population.

The Texas Cancer Information Center (Texas Cancer Registry), formerly known as the Texas Cancer Data Center tracks mortality on the different types of cancer in Texas. From this data center, Texas lung cancer mortality for selected years by age, sex,

race, and geographic area can be obtained. As of now, Texas is not included in SEER. One major difference between the Texas Registry and SEER is that the race Hispanic is defined as a separate race class in the Texas Cancer Registry. This resource will also be providing cancer incidence information in the near future.

As seen in the graphs on the next page, overall Texas and SEER mortality rates are similar in both men and women.

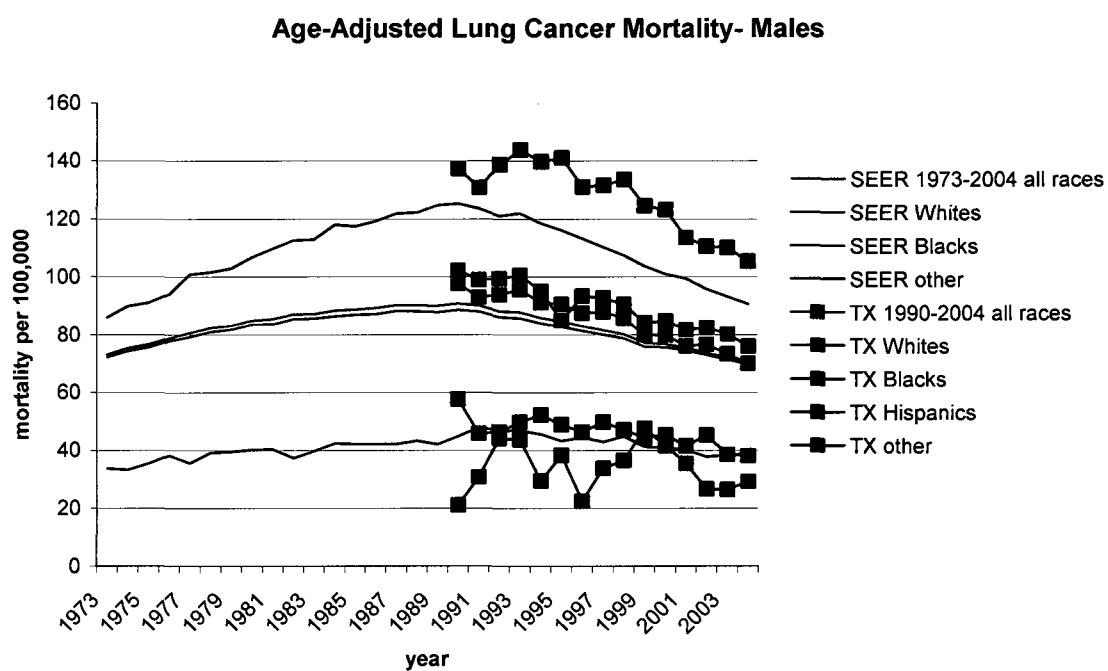


Figure 1.1 Age-adjusted lung cancer mortality rates for males

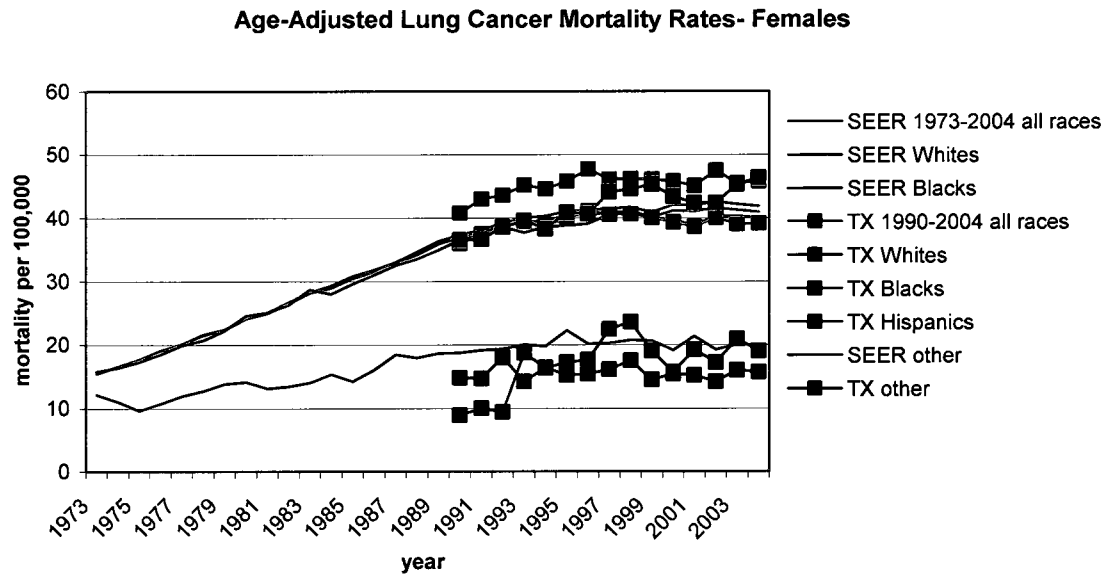


Figure 1.2 Age-adjusted lung cancer mortality rates for females

Although data is collected on Hispanics, data on Hispanic heritage is collected separately from race. So, Hispanics can be included in any of the race assignments, whites, blacks, and other. Other races in SEER include American Indians, Alaska Natives, and Pacific Islanders.

1.4 Risk Factors for Lung Cancer

The primary risk factor for lung cancer is smoking but there are other known risk factors including exposure to asbestos, radon, and secondhand smoke. There is also known to be a genetic risk effect because relatives of individuals with lung cancer have a higher risk of developing lung cancer themselves.

1.4.1 Smoking Tobacco

Tobacco has been used for centuries all over the world. However, following the introduction of manufactured cigarettes with addictive properties the incidence of lung cancer has risen rapidly. Scientist in Nazi-era Germany carried out the first studies on the relationship between smoking and lung cancer (Proctor 1999). By the 1950's case-control trials in both the United States and Britain showed a strong association between cigarette smoking and lung cancer (Doll and Hill 1950, Levin et al. 1950, Wynder and Graham 1950). The causal relationship between smoking and lung cancer was confirmed by large cohort studies including the British Physicians' Study, and the American cancer society's Cancer Prevention Study (US DHEW 1964). As lung cancer incidence is tracked over time it has been shown to follow cigarette consumption trends (Wingo et al. 1999). Cigar and pipe smoking has also been shown to be a risk factor for lung cancer, but on a much lower scale than cigarette smoking (Boffetta et al. 1999). The lower risk associated with smoking tobacco in these forms can be explained by the increased amount of carcinogenic additives in manufactured cigarettes (Alberg and Samet 2003) and by inhaling cigarettes' smoke deep into the lung as opposed to just the mouth and throat when smoking pipes and cigars (Alberg and Samet 2003).

1.4.2 Secondhand Smoke

Exposure to secondhand smoke (SHS) has been shown to be a risk factor for lung cancer (Brennan et al. 2004, Gorlova et al. 2006, Vineis et al. 2005, Surgeon General's

Report 2001 and 1986, NIH 1999, National Research Council (NRC) 1986, EPA 1992, IARC 2002).

An EPA report published in 1992 indicated that SHS was causally associated with increased lung cancer risk (Brown 1992) resulting in an additional 3,000 deaths from lung cancer per year. This was determined through a meta-analysis of 30 spousal exposure studies where never-smokers that lived with smokers were compared with those that did not. Further studies concluded that there was an increased risk of 20-25% of LC for never smokers that were exposed to SHS through their spouse or workplace.

1.4.3 Environmental Exposures

Workplace or residential exposures such as radon, asbestos, arsenic, silica, and chromium have also been shown to increase risk of LC (International Agency for Research on Cancer 1986, Alberg et al. 2005, Gottschall 2002, Neuberger and Field 2003). Many of the workplace exposure risk factors have been shown to have a synergistic effect when combined with smoking, meaning that the risk for smokers is increased more than additively for each risk factor.

Radon is a radioactive gas that is produced when uranium from rocks and soil decays. Underground mines contain high levels of radon gas. Studies of uranium miners (Lubin et al. 1994) and animal studies (Cross 1994) have established a causal relationship between radon gas and LC. Although radon is a major exposure for uranium miners, it has also been found to be present in people's homes.

Asbestos is a substance made of long thin fibers used by manufacturers in the late 19th century due to its resistance to heat, electricity and chemical damage. Inhalation of

asbestos fibers has been causally linked to many respiratory illnesses including LC (Selikoff et al. 1964).

There is some belief that air pollution may cause lung cancer because of all the known carcinogens that are in the air. Taking into account the large daily human consumption of air, even low levels of carcinogens can be a health concern. Extrapolation about the known risks of lung cancer by the occupational exposures gives evidence to the conclusion that air pollution may be a risk factor for lung cancer (Doll and Peto 1981, Friberg and Cederlof 1978, Doll 1978).

1.4.4 Pre-existing Lung Disease

There are two main categories of lung disease. The first includes disorders that obstruct airflow such as chronic obstructive pulmonary disease (COPD), and in its most severe form, emphysema. The second group includes disorders where lung capacity is restricted due to inhaled fibrous substances as in black lung disease. Having obstructive lung disease such as COPD has been shown to increase risk for lung cancer in numerous studies (Littman et al. 2004, Mayne et al. 1999, Tockman 1994, Wu et al 1995).

1.4.5 Family History of Cancer

Although smoking is the single most important risk factor for lung cancer, genetic factors also play a role in lung cancer development. After accounting for smoking, several studies have found an increased risk in relatives of lung cancer cases compared to relatives of lung cancer controls (Sellers et al. 1987, Schwartz et al. 1996, Schwartz et al.

1999; Etzel et al. 2003, Cote et al. 2005). It has also been shown that 1st degree relatives (parents, siblings, and offspring) of never smokers with lung cancer have an increased risk of LC (Gorlova et al. 2006, Wu et al. 1996, Brownson et al. 1997).

A recent analysis of 11 studies on never smokers concluded that there was a 1.5 fold increase of LC for those with family history of LC (Matakidou et al. 2005). This result has also been shown for the relatives of smokers previously diagnosed with lung cancer (Etzel et al. 2003, Broman et al. 2000, Mayne et al. 1999).

1.4.6 Genetic Susceptibility of lung cancer, DNA repair capacity

In a recent linkage analysis, a lung cancer susceptibility region 6q23-25 was identified (Bailey-Wilson et al. 2004). A more recent study found evidence of association of lung cancer with SNPs located in 15q region spanning CHRNA3, CHRNA5, and PSMA4 genes (Amos et al. 2007). Also, molecular epidemiological studies have documented a substantially increased risk for lung cancer associated with poor DNA repair capacity following exposure to mutagens (Wei et al. 2000, Shen et al. 2003). In these studies DNA repair capacity was evaluated by a host-cell reactivation assay that measures cellular ability to remove adducts from plasmids transfected into lymphocyte cultures *in vitro*, by expression of damaged reporter genes. A dose-response relationship has been demonstrated with increasing risk associated with poorer repair capacity (Shen et al. 2003). DRC is modulated by polymorphisms in genes in the nucleotide excision repair pathway (Qiao et al. 2002), which suggest its genetic determination. However, not only genetic factors but also smoking tends to modulate DRC: It is highest in current

smokers, followed by former and never smokers (Shen et al. 2003). This can be explained by the cells' need for increased repair in the presence of carcinogen/mutagen exposure.

1.5 Attributable Risks

Attributable risks are studied to determine the estimated percentage of lung cancer cases caused by certain risk factors. Below is a table of the estimated attributable risks for some of the risk factors mentioned previously. Data is taken from Alberg and Samet 2003.

Table 1.1 Attributable risk to lung cancer in the US population

Risk Factor	Attributable Risk
Smoking	90%
Occupation exposures	9 to 15%
Radon	10%
Outdoor air pollution	1 to 2%

1.6 Gender Differences in Lung Cancer

Lung cancer may differ in women and men with respect to several characteristics not necessarily caused by gender differences in smoking habits. Whether there is a gender difference in lung cancer susceptibility remains a controversial issue. Some case-control studies concluded that the risk of developing lung cancer due to smoking is higher for women than for men (Risch et al. 1993, Zang and Wynder 1996), whereas other case-control as well as cohort studies show that the risks are similar (Bain et al. 2004, Prescott et al. 1998, Sobue et al. 2002).

Among never smokers, it has been shown that women are disproportionately more affected by lung cancer than men (Subramanian et al. 2007). Even after adjusting for the fact that there is a larger proportion of never smoking females than never smoking males, there is still a small increased risk for female never smokers as compared to male never smokers. Wakelee et al. (2007) in the analysis of several cohorts have shown that lung cancer incidence in never smoking women was higher than in never smoking men for most of the cohorts analyzed.

A recent finding by IELCAP investigators Henschke and Miettinen (2006) showed that females, with tobacco exposure similar to that in men, had a higher risk to be diagnosed with lung cancer on CT screen, indicating that women are more susceptible to tobacco carcinogens. However, an analysis by Bain et al. (2004) comparing the Nurses Health Study and the Health Professionals Study showed that there were no lung cancer incidence differences for men and women with similar smoking histories. This study also noted a non-statistically significant increase in the lung cancer incidence for female smokers above the age of 60 compared to male smokers.

The progression rate of the disease also shows gender differences. It was suggested that women have slower growing tumors than men (Hasegawa et al. 2002, Usuda et al. 1994,). SEER data suggest that females tend to be diagnosed with a less advanced stage of lung cancer compared to men. However, the age at diagnosis over all cell types combined is slightly earlier in women than in men. SEER data show that postmenopausal women (older than 50) had better lung cancer survival than men in the same age group while there was no difference for younger females vs. males (Moore et al. 2003). To what extent these differences are caused by differences in smoking patterns

between females and males, which are substantial (Patel et al. 2004), is not clear because SEER data on smoking histories is not available. A TSCE model fitted to the gender-specific data from the MDACC case-control study, the Nurses Health Study, the Health Professional Study and CPS-I, all of which provide information on smoking, will help answer these questions related to the gender differences in lung cancer progression at its early stages.

1.7 Lung Cancer in Never Smokers

Many differences between lung cancers in smokers compared to lung cancers in never smokers have been shown (Sun et al. 2007). First, as noted in the previous section women never smokers are more often affected than men. Also, there is an increased incidence of the adenocarcinoma histological type of lung cancer in never smokers. Studies from Asia including data from Japan, Hong Kong, and Singapore show that there is an earlier age at diagnosis for never smokers with lung cancer than smokers with lung cancer (Toh et al. 2006, Shimizu et al. 1984, Koo et al. 1985). However, many studies from the United States and Europe show that the age at diagnosis for never smokers with lung cancer is the same or older than that for current smokers (Dibble et al. 2005, Brownson et al. 1998, Muscat and Wynder 1995, Wakelee et al. 2007, Nordquist et al. 2006). The Wakelee study included data on 9 gender specific never smokers cohorts all of which showed a same or later age at onset for never smokers lung cancer as seen in the following table.

Table 1.1 Median age at LC diagnosis by smoking status (Wakelee et al. 2007)

Cohort	Smoking Status		
	Never	Former	Current
Nurses Health Study, female	64	68	64
Health Professionals' Follow-up Study, male	67	71	68
California Teachers' Study, female	67	70	67
Multiethnic Cohort Study, male	72	72	69
Multiethnic Cohort Study, female	72	70	67
Swedish Uppsala/Orebro Lung Cancer Registry, male	64	71	64
Swedish Uppsala/Orebro Lung Cancer Registry, female	67	66	63
First National Health and Nutrition Examination Survey Epidemiological Follow-up Study, male	78	72	69
First National Health and Nutrition Examination Survey Epidemiological Follow-up Study, female	71	67	62

The lower ages at diagnosis seen in the Asia studies may be explained by more exposures to other risk factors including exposure to poorly ventilated fumes from cooking and the burning of coal. Studies have shown that cooking oil fumes, common in Asia, are a risk factor for lung cancer (Boffetta and Nyberg 2003).

1.8 Screening for Lung Cancer

Screening for lung cancer is currently a very controversial topic. There are 3 widely cited studies from the 1970's that all failed to show a benefit from x-ray screening for lung cancer (Early lung cancer detection 1984). However, advances in technology specifically CT, gives some hope for a screening tool for lung cancer. The 5-yr survival for rate for stage I lung cancer is 70% while the 5-yr survival rate for stage IV is only 5% (Unger 2006). If lung cancer can be diagnosed earlier when survival is better, then there may be hope in lowering lung cancer mortality.

Some studies indicate a possible benefit for CT screening in Lung Cancer (IELCAP 2006) in terms of increased survival for stage I screen-detected lung cancers. However, this study is missing a control arm for comparison, and other studies dispute any significant mortality reduction (Bach 2007).

1.8.1 CT Screening for LC

The first published American study on CT screening for lung cancer was the Early Lung Cancer Action Program (Henschke et al. 1999, Henschke et al. 2001). IELCAP recently reported among participants with screened detected stage I lung cancer that underwent resection within 1 month of diagnosis the 10-year survival was 92% (IELCAP 2006) while 8 patients who declined surgery all died within the next 5 years. The authors argue that screening detects lung cancer at an earlier stage where curability is

higher leading to more lives saved. However, there is no control arm in this study with which to compare survival and mortality.

Also, a measure such as survival in this type of study can be biased (Aberle 2008). One type of bias is lead-time bias. This type of bias suggests that survival is increased simply because the cancer is diagnosed earlier but the diagnosis did not delay death from lung cancer at all. In essence, the individual discovered they were going to die sooner than they would have without screening, but treatment did not do anything to change the overall outcome. So, the time between lung cancer diagnosis and death was increased just because the cancer was diagnosed earlier and there was no effect on delaying the death. Another possible bias is length bias where screening is more likely to pick up slowly progressing cancers than faster growing cancers. The third bias which is the extreme case of length bias, is termed “overdiagnosis”. The overdiagnosis argument is that screening mostly detects very slow-growing cancers that would not have progressed anyway leading to “overdiagnosis” of lung cancer in the screened group. Comparing the mortality in screened arm to that in a control arm is the only way to adjust for these possible biases.

A recent study by Bach et al. used a previously validated model for lung cancer risk (Bach et al. 2004, Cronin et al. 2006) in order to simulate a control arm for 3 small CT screening trials. Using this method it was found that there was no reduction in mortality between the observed screened arm and the expected mortality from the model (Bach et al 2007). One major criticism of this study is that it is sensitive to the assumptions used in order to fit the model, and the fact that the follow-up time in the screening studies was too short.

These two studies with opposing viewpoints (IELCAP 2006, and Bach 2007) were published within 6 months of each other. A nice discussion comparing the 2 studies and their strengths and weakness was written by Black and Baron (2007). In this paper Black and Baron state that the best way to avoid the problems associated with these 2 studies is using a randomized control trial (RCT) to determine if lung cancer screening is effective. There is an ongoing RCT of screening for lung cancer comparing CT screen with x-ray screen being conducted in the United States, National Lung Screening Trial (NLST) but there are concerns that it may not be powered to detect a difference in mortality because the control group is not receiving standard of care (no recommended screening) but is in fact being screened as well using chest x-ray.

Chapter 2

Carcinogenesis Modeling

2.1 Review of Carcinogenesis Models

In the 1950's Armitage and Doll (1954) introduced a multi-stage model to describe carcinogenesis. The Armitage-Doll model described the process of carcinogenesis as a finite number of mutations turning a normal cell into a malignant cell. Normal cells undergo k distinct transformations to become malignant. This model was developed through the observations that for many cancers, the log of incidence was linear in log of age. Incidence was modeled as: $I(t) = a \times t^{k-1}$, where t is age. This was derived from the following. The probability that mutation i occurs in $[0, t]$ is $p_i t$. Then by time t ,

the probability that $k-1$, mutations occurred is $\prod_{i=1}^{k-1} (p_i t)$. There are $(k-1)!$ possible

combinations of the mutations, only one of which is in the right order to produce the

malignant transformation. This gives the density function $f(t) = \frac{1}{(k-1)!} \prod_{i=1}^{k-1} (p_i) t^{k-1}$,

which leads to an incidence function of the form: $I(t) = a \times t^{k-1}$.

In their 1954 paper, the model was successfully fit to English data on the incidence of esophageal, stomach, pancreas, and colon cancers. However, the model did

not seem to fit incidence data on lung, bladder, prostate, and female reproductive cancers. The authors suggested that these cancers did not fit because of the involvement of other risk factors such as smoking, and hormonal fluctuations.

2.1.1 Two-stage Carcinogenesis Models

It was then theorized that the poor fit could result from the fact that the Armitage-Doll model ignores the fact that many cancers develop from the proliferation of pre-malignant abnormal cells. In response to this Armitage and Doll (1957) introduced a model with two stages, the first step involving the mutation of a normal cell (NC) into an intermediate cell (IC), and the second step involving the further mutation of the intermediate cell into a malignant cell (MC). This newer model allowed for the proliferation of the intermediate cells by modeling it as exponential growth of the intermediate cells in the first stage and malignant cells escape from control in the second stage.

In 1960, Kendall introduced a two-stage model in which 1st stage involved a subcritical birth and death process (birth rate < death rate) of the intermediate cells. The second stage involved a supercritical birth and death process of malignant cells. Neyman and Scott (1967) introduced a two-stage model where the IC's generate clones according to a subcritical birth and death process.

2.2 Two-stage clonal expansion (TSCE) Model

With the advance of molecular biology, clonal expansion was recognized as an essential stage in carcinogenesis (Marks et al. 2007, Sikkink et al. 2003). Motivated by the idea of clonal expansion, Moolgavkar et al. (1979) established a two-stage clonal expansion (TSCE) model. This model is depicted as follows:

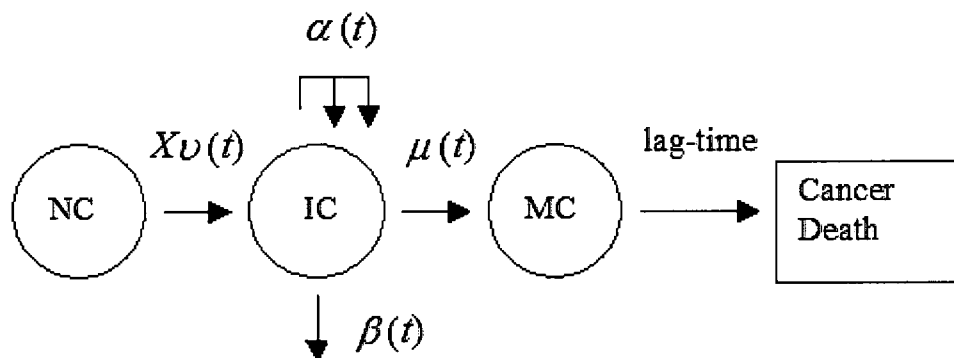


Figure 2.1 Depiction of the TSCE model

The TSCE model assumes that a normal cell (NC) mutates into an initiated cell (IC) in the first transition, according to a Poisson process with intensity $\nu(t)$, where t denotes the age. There are X normal cells in the tissue at birth or maturity, depending on the tissue. Then the initiated cell duplicates or dies according to a birth-death process with parameters $\alpha(t)$ and $\beta(t)$ and forms a clone of initiated cells. Each initiated cell can also mutate into a malignant cell (MC) for the second transition according to a Poisson process with parameter $\mu(t)$. After a lag time, this malignant cell is assumed to develop into a cancerous tumor with probability one. The distribution of the lag-time can be specified and the parameters of the lag-time distribution can be fitted simultaneously with the parameters of the TSCE model. This model also allows for the growth of the number of NC's (X) through a deterministic model.

This modeling framework allows for evaluating the effects of different lung cancer risk factors on the model parameters and ultimately on cancer incidence, by making the model parameters dependent on the risk factors through response functions.

2.2.1 Non-identifiability in the TSCE Model

The TSCE model is an incidence-based model, meaning that it is set up to fit incidence data. One deficiency with the model is that only $4k-1$ of the $4k$ biological parameters ν , μ , α , and β are identifiable when fitting to data in the piecewise-constant parameters over k distinct time intervals. There are two widely accepted approaches to dealing with the non-identifiability problem. The first approach is to set the background mutations rates equal to each other, $\nu_0 = \mu_0$, and assume a reasonable number of normal cells such as, $X = 10^7$. This approach makes use of the fact that the only the product ($\nu\mu$) appears in the survival and hazard functions. So, using this assumption will not affect estimates of incidence rates and risk.

Another possible approach to the non-identifiability problem is to use a new set of parameters. Heidenreich et al. (1997) found that the following 4 parameters are identifiable when fitting the TSCE model to incidence data in the piecewise constant case:

$$y_i = \nu_i \mu_0$$

$$m_i = \mu_i / \mu_0$$

$$\gamma_i = \alpha_i - \beta_i - \mu_i$$

$$q_i = (-\gamma_i + \sqrt{\gamma_i^2 + 4\alpha_i \mu_i}) / 2$$

The parameters, Y , m , γ , and q have biological interpretations, yet they are less straightforward than the original biological based ones. Y can be interpreted as the initiation rate, m as the malignant transformation rate, and γ as the net proliferation rate of the ICs. The q parameter is related to the asymptotic height of the hazard function but is more difficult to interpret. All of these parameters are identifiable when the model is fitted to incidence data.

2.2.2 Evolution of the TSCE model

As mentioned in the previous section, Moolgavkar et al introduced the TSCE model in 1979. Since then there have been many discoveries regarding this model. In the early 1990s Tan derived a general formula to calculate the hazard and survival functions of the time to first malignant cell under the TSCE model with piece-wise constant parameters (Tan 1991). These formulas were derived using probability generating functions but required a numerical approximation to solve partial differential equations.

In 1994 the closed form solutions to the hazard and survival function of time to the first malignant cell were derived independently by Kopp-Schneider et al. (1994) and Zheng (1994) under the constant parameter setting. Then in 1996 Heidenreich showed that only three of the four parameters were identifiable when fitting to incidence data and suggested alternative identifiable parameters.

Finally in 1997, Heidenreich was able to derive the exact hazard and survival functions for the TSCE model under piecewise constant parameters. This paper included easily programmable recursion formulas for the exact calculations and made it easier to use the model because it did not require numerical approximations. In this paper,

Heidenreich argued that in the piecewise constant setting over k subintervals of time, only $4k-1$ parameters out of $4k$ are identifiable. This was mathematically proved in Deng's thesis (Deng 2006).

2.2.3 TSCE Models Applied to Lung Cancer

The TSCE model is widely used in risk analysis and has been used before in the estimation of the effects of smoking and other risk factors on transition rates and incidence of lung cancer.

In 1990 Moolgavkar and Luebeck fit the TSCE model with time dependent parameters to data on radon induced lung tumors in rats. This model required the numerical approximation to calculate the survival and hazard functions. Later Heidenreich et al. (1999) fit the same data to the TSCE model using the exact formulas. This analysis also differed from the previous in that it defined two types of lung tumors in the radon exposed rats, incidental and fatal.

The model has been fit to many cohorts to estimate the effects of different risk factors on lung carcinogenesis. In 1997 atomic bomb survivors were used to estimate the risk of large exposures to radiation (Heidenreich et al. 1997). Data on Colorado uranium miners were modeled to estimate the effects of smoking and radon exposure on lung cancer risk (Luebeck et al. 1999). A cohort of Chinese tin miners was used to estimate the effects of tobacco, arsenic, and radon on the incidence of lung cancer (Hazelton et al. 2001). In 2006 a Canadian cohort was fit to estimate the effects of whole body radiation on lung carcinogenesis (Hazelton et al. 2006). More recently, the model has been applied to cohorts from CPS-I, CPS-II, the British Doctors Study, the Health Professionals

Follow-up Study, and the Nurses Health Study to estimate the effects of smoking on lung carcinogenesis (Hazelton et al. 2005, Meza et al. 2007).

2.3 Fitting the TSCE Model

In this section we outline the traditional maximum likelihood approach used to fit the TSCE model to prospective cohort data. Then we introduce a new resampling based method of reconstructing cohort data from the combination of case-control data on risk factor dependencies and tabled incidence/mortality rate data. This method is then applied to a simple model.

2.3.1 Prospective Cohort-based Likelihood

The TSCE model is usually fit to prospective cohort data using maximum likelihood. The cohort likelihood is defined as the product of the individual likelihoods,

$L = \prod_j L_j$. For a fixed lag-time t_{lag} , each L_j depends on the time of entry into the study, s_j

, censoring or failure time, t_j , and the individual's exposure history.

$$L_j(t_j, s_j) = \begin{cases} h(t_j - t_{lag})S(t_j - t_{lag})/S(s_j - t_{lag}) & \text{if diagnosed with cancer} \\ S(t_j - t_{lag})/S(s_j - t_{lag}) & \text{otherwise} \end{cases}$$

2.3.2 Re-sampling Based Method of Reconstructing Cohort Data

In order to fit the TSCE model to case-control data a new method was developed to reconstruct cohort data using the combination of case-control data and tabled

incidence/mortality data using re-sampling. The goal of the method is to re-sample cases and controls in proportions reflected in the mortality data to recreate cohort data. The main assumption of this method is that given all matching stratam, cases and controls represent a random sample from the population. Each re-sampled cohort is referred to as a pseudo-cohort and is created by sampling individuals. Each individual is sampled as follows:

1. For each individual, the age bin (5year) in which the person belongs is sampled based on the number of individuals in each age bin of the case-control study.
2. Based on the age bin generated above, randomly sample whether the individual gets cancer or not based on the estimated probability of an individual within the sampled age bin getting cancer within the 5 years spanning the age bin.
3. Once we have an age bin, and cancer status we then sample an individual from the case-control dataset with the same characteristics and use information on their risk factor exposures.
4. The censoring or failure time of the individual is assigned as their age at enrollment from the case-control dataset and the age at entry is assigned as 5 years prior for controls and at a randomly distributed age in the previous 5 years for cases.

Ages of enrollment and exit were done this way because the cancer status was sampled from the probability of getting cancer over a 5-year interval. If the individual does get cancer during the interval the timing is sampled as uniform over the interval.

25,000 individuals are re-sampled from the case-control dataset for each pseudo-cohort created. Then each pseudo-cohort is fit to the TSCE model by maximizing the cohort likelihood in the usual way. Two hundred pseudo-cohorts are created and fitted. This provides 200 joint estimates of the parameters for each simulated case-control trial. The overall fit is assumed to be the mean estimates over the 200 runs.

2.3.3 Re-sampling Method Applied to a Simple Model

To further explore this method, a simpler model was assumed. For this simpler model, individuals go from normal to lung cancer as a simple poisson process with parameter, λ . Under this model, we can look at the likelihood in terms of only one parameter.

Under the exponential model, case-control data was simulated as outlined in the next section. The parameter of the exponential, λ , was set as 1.5×10^{-4} to approximate the lifetime risk of cancer as 1.5% which is comparable to the lifetime risk of lung cancer in never smokers. Using this new method outlined above the following negative log-likelihood was calculated for a simulated reconstructed cohort using the simple exponential model.

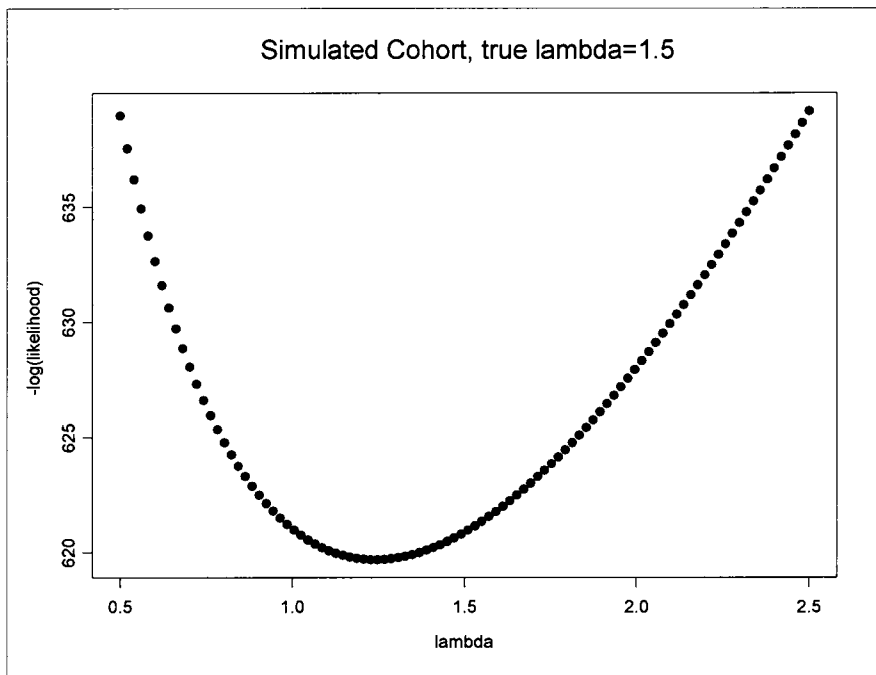


Figure 2.2 Cohort likelihood applied to reconstructed pseudo-cohort data

Using the fitting regime outlined the following results were obtained for fitting the simple exponential model:

Table 2.1 Results of fitting the exponential model using the resampling method

parameter	true value	mean fit	LCL 95%	UCL 95%
$\lambda \times 10^4$	1.5	1.507	0.978	1.943

The accurate results for fitting the simple exponential model using the new re-sampling method indicated that the method has promise for fitting the TSCE model to case-control data and prompted further simulation studies.

Chapter 3

Simulation Studies

Simulation studies were conducted on the accuracy of fitting the TSCE model using the resampling based method of reconstructing time to event data from the combination of case-control data and incidence/mortality rate data. This was compared to the accuracy in fitting traditional prospective cohort time to event data.

3.1 Simulation Methods

Simulation studies to examine the efficacy in fitting the background rates (in the absence of risk factors) of the TSCE model to simulated data were carried out using the 2 different parameterizations to both traditional cohort data and using the new re-sampling method on case-control data. The background rates were used because they are often the hardest to estimate particularly in the case-control setting.

First, individuals were simulated using the following simulation routine suggested by Kaiser and Heidenreich (2004).

1. For each individual a death of other causes time, t_d , distributed as $N(80, 15^2)$ was simulated.

2. Then the probability of not developing cancer by age t_d was calculated according to the TSCE model: $p_{TSCE} = S(t_d)$.
3. Then a uniform(0,1) random variable, u , was drawn.
 - a. If $u \leq p_{TSCE}$ then time of censoring is t_d and no cancer develops during the individual's lifetime.
 - b. If $u > p_{TSCE}$ then cancer develops during the individual's lifetime and is diagnosed at age, t , computed by inverting the survival function, $u = S(t)$.

Simulating 1,000,000 individuals creates a population. The population is then truncated to only include individuals aged 30-84. The tabled age-specific (5 year bins) incidence rates are then calculated from the simulated population for use in the resampling method. The simulation of cohorts is done by randomly sampling 25,000 individuals from the simulated population. For simplicity, the age at entry to the study is assumed to be birth for all individuals, and the lag-time from birth of the first malignant cell to lung cancer is assumed to be zero.

The simulation of each case-control study is done by randomly sampling cancer cases from the simulated population with ages of diagnosis in the range 30-80, comparable to entry criteria. Then an age at enrollment is simulated as $N(55,10^2)$ for the controls (all the individuals in the simulated population that did not get cancer). This age at enrollment is reflective of the ages of unmatched controls in the MD Anderson case-control data. The controls are then sampled from the population with ages of enrollment within 5yrs of the ages at diagnoses of the cases sampled previously. Each simulated case-control study consists of 1,500 cases and 1,500 controls.

3.2 Simulated TSCE Parameters

The following sets of parameters were used to simulate the populations based on the 2 methods of dealing with non-identifiability. The values chosen come from two studies fitting the TSCE model. The biological parameters come from a study by Hazelton that fit to the males of the CPS-I study and the identifiable parameters come from a study by Heidenreich fitting males in a large German case-control study.

Biological parameters with constraints (Hazelton 2005):

$$\alpha_0 = 22.65$$

$$\gamma_0 = \alpha_0 - \beta_0 - \mu_0 = 0.075$$

$$\mu_0 = \nu_0 = 1.4 \times 10^{-7}$$

$$X = 10^7$$

Identifiable parameters (Heidenreich 2002):

$$y_0 = \nu_0 \mu_0 = 0.11 \times 10^{-7}$$

$$g_0 = \alpha_0 - \beta_0 - \mu_0 = 0.134$$

$$q_0 = (-g_0 + \sqrt{g_0^2 + 4\alpha_0\mu_0})/2$$

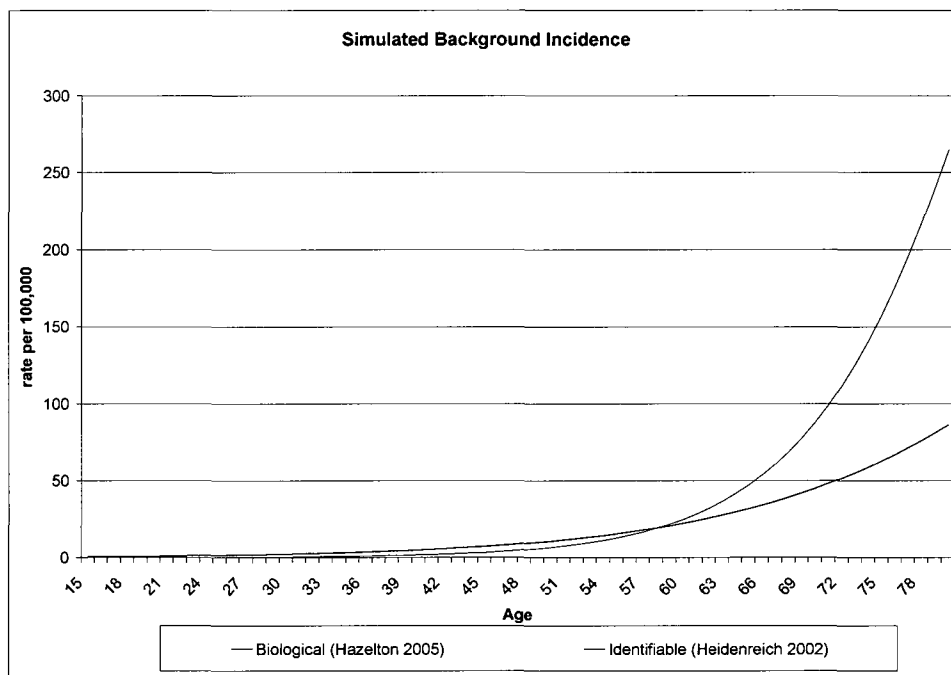


Figure 3.1 Simulated age-specific incidence rates

3.3 Results for Traditional Prospective Cohort Method

As described earlier, cohort data were simulated according to the TSCE model and then parameters were estimated using maximum likelihood. The following table shows the results of the simulation study.

Table 3.1 Results of fitting simulated data using traditional prospective cohort method

Type	parameters	true value	Mean fit	LCL 95%	UCL 95%
Biological	α	22.65	23.400	7.327	40.499
	γ	0.075	0.0759	0.0660	0.0857
	$\mu \times 10^7$	1.4	1.3784	1.0541	1.7233
Identifiable	$y \times 10^7$	0.11	0.1125	0.0603	0.1788
	g	0.134	0.1343	0.1257	0.1443
	$q \times 10^6$	1.2	1.2128	0.7255	1.8943

The following plots contain the actual and average predicted incidence from the 200 simulated studies as well as predicted incidence for 20 random runs.

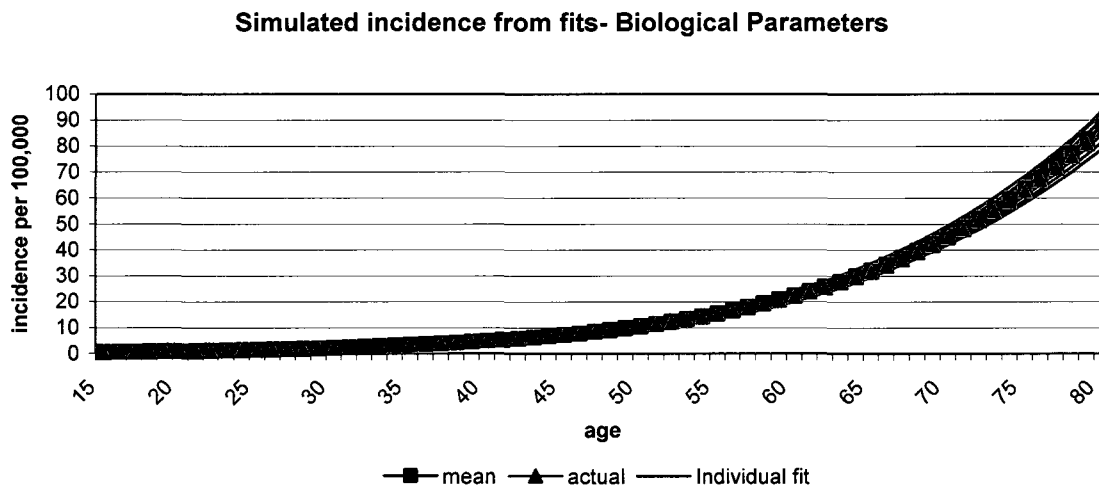


Figure 3.2 Predicted incidence rates- traditional cohort method with biological parameters

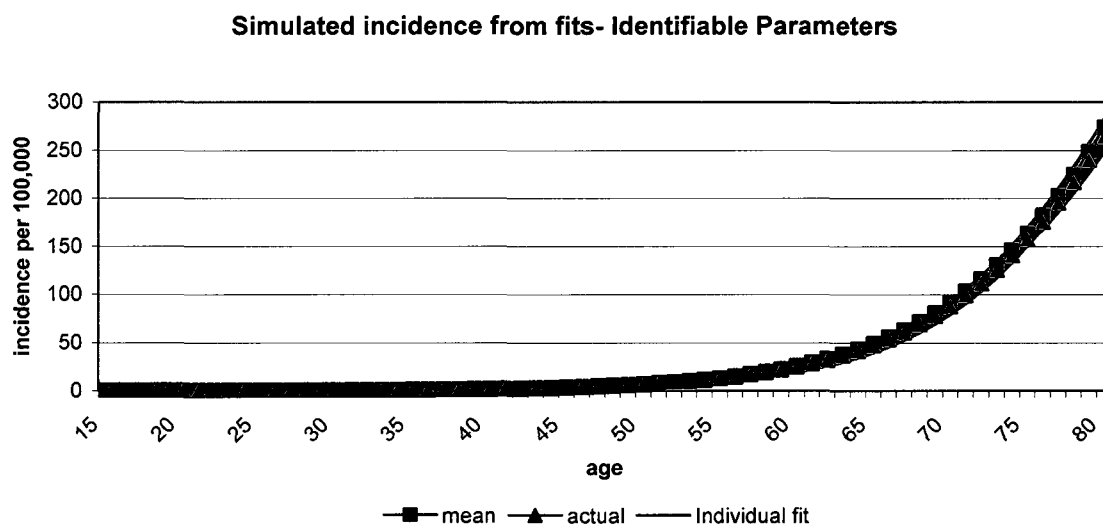


Figure 3.3 Predicted incidence rates - traditional cohort method with identifiable parameters

The top graph is for the biological parameters while the lower graph is for the identifiable parameters. Both sets of parameters can be accurately fit using the traditional cohort likelihood as expected.

3.4 Results for Resampling Method

As mentioned previously, a simulation study was conducted on the effectiveness of reconstructing cohort data using resampling to fit the TSCE model. This method seeks to use case-control data on risk factors and mortality rate data to reconstruct time-to-event data as would be seen in a prospective cohort study. The following table shows the simulated fits as compared to the actual parameters.

Table 3.2 Results of fitting simulated data using the resampling method

Type	parameters	True value	mean fit	LCL 95%	UCL 95%
Biological	α	22.65	25.271	$\sim 10^{-4}$	65.487
	γ	0.075	0.0787	0.0612	0.1032
	$\mu \times 10^7$	1.4	1.3301	0.7028	2.0042
Identifiable	$y \times 10^7$	0.11	0.1086	0.0189	0.2630
	g	0.134	0.1372	0.1186	0.1600
	$q \times 10^6$	1.2	1.1462	0.2888	2.7148

The method appears to be accurate in fitting the baseline parameters of the TSCE model but not as precise as using traditional cohort data resulting in wider confidence intervals. The overall procedure was then repeated 200 times and average parameter fits were calculated. The results are depicted in the following graphs.

Simulated incidence from fits- Biological Parameters

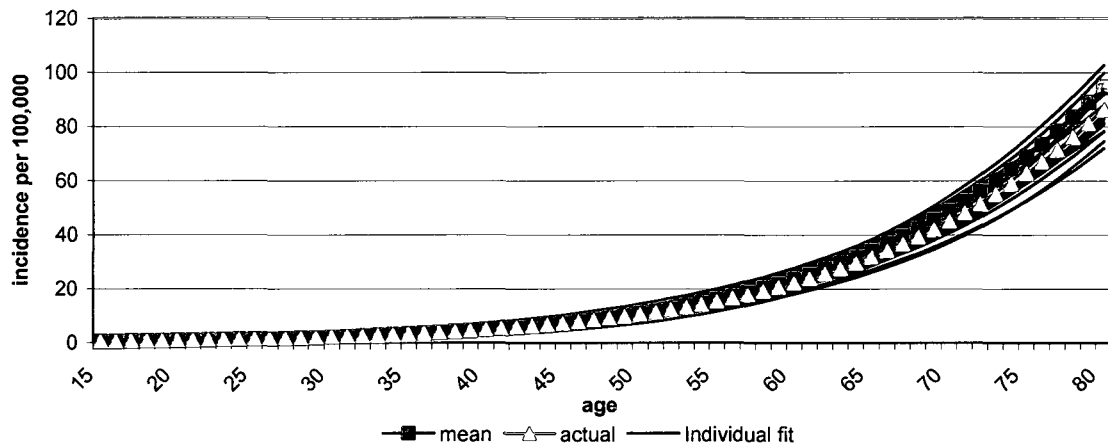


Figure 3.4 Predicted incidence rates- resampling method with biological parameters

Simulated incidence from fits- Identifiable Parameters

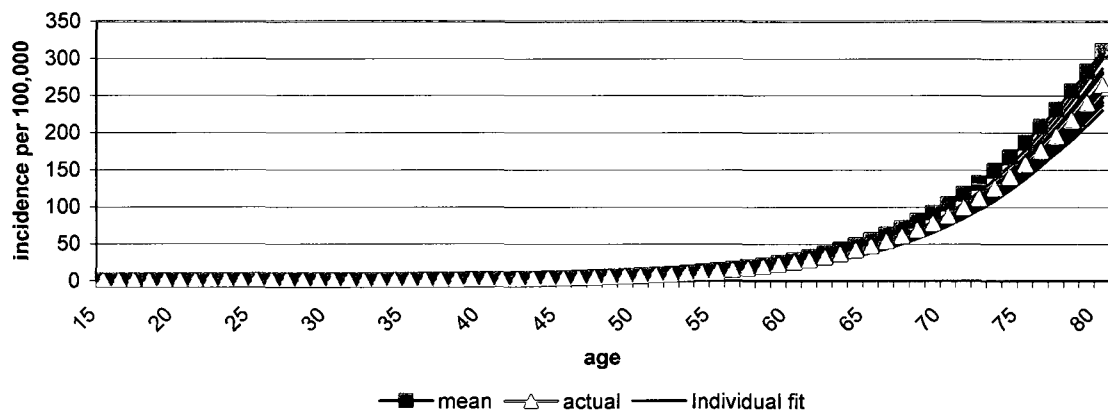


Figure 3.5 Predicted incidence rates- resampling method with identifiable parameters

The mean incidence rate was calculated as the average parameter fit over the 200 simulations. The identifiable parameters seem to produce some bias when averaging. Therefore, we chose the biological parameters for fitting the model to the MDA case-

control data because they provide a better fit and are easily interpretable.

Chapter 4

Smoking Based TSCE Model and Validation

A smoking based TSCE model was fit using the resampling based method. A model based of smoking alone was chosen so that it could be used in CISNET's Smoking Base Case, as well as, to simulate LC mortality in the absence of screening for a single arm CT screening trial. The data used to fit the model, the method of reconstructing cohort data, and the final fitted model are presented in this chapter. Also, simulation is used to validate the model as a predictor of lung cancer mortality in the non-asbestos exposed control arm of CARET. The model is also used as part of the CISNET Smoking Base Case Project.

4.1 MD Anderson Case-Control Data

An important component to this project is to determine the effect of different risk factors on the process of cancer development. Datasets such as SEER contain information about overall lung cancer incidence and mortality but do not provide information about individual risk factors such as smoking. A case-control study is currently underway in the M.D. Anderson Cancer Center Department of Epidemiology under the direction of Dr. Spitz. In this study, measurements of DNA repair capacity, as well as data on other risk factors such as smoking are being recorded. Cases of lung cancer are matched with

cancer-free controls on age (within 5yrs), gender, ethnicity, and smoking status. The MD Anderson case-control data contains information on over 6,000 matched cases and controls.

Table 4.1 Characteristics of cases (n=3433) and controls (n=3132) available from the lung cancer case-control study (R01 CA55769, Spitz, PI)

Characteristic	Cases	Controls
Mean Age (SD)	62.3 (11.1)	59.4 (10.9)
Race/Ethnicity	n (%)	n (%)
White	2744 (79.9)	2488 (79.4)
Black	488 (14.2)	411 (13.1)
Hispanic	173 (5.0)	210 (6.7)
Other	28 (0.8)	23 (0.7)
Sex		
Male	1848 (53.8)	1587 (50.7)
Female	1585 (46.2)	1545 (49.3)
Smoking Status		
Current	1371 (39.9)	1120 (35.8)
Former	1453 (42.3)	1274 (40.7)
Never	556 (16.2)	702 (22.4)

Analyses in this study will be based on a subset of whites, 992 males and 919 females, because of the low sample sizes of the other races. Even though cases and controls were matched on smoking status, the cases smoked more than the controls as seen in terms of average pack-years.

Table 4.2 Pack-year histories for 992 white males and 919 white females included in analysis

Characteristics		Smoking Status	Number	Avg Pack Yrs
Males	Cases	Current	214	58.7
		Former	259	53.1
		Never	28	0.0
		Total	501	55.3
	Controls	Current	168	54.5
		Former	287	43.4
		Never	36	0.0
		Total	491	44.0
Females	Cases	Current	211	50.1
		Former	183	41.1
		Never	58	0.0
		Total	452	40.0
	Controls	Current	181	36.7
		Former	197	36.2
		Never	99	0.0
		Total	467	29.7

A family history analysis was done to determine the effects of family history on lung cancer risk. Logistic regression analysis was performed to test the association of cancer family history with lung cancer risk. Three definitions of a positive family history were considered. The first definition was 2 or more first degree relatives diagnosed with any cancer (FH1). The second was at least one relative diagnosed with lung cancer (FH2) and finally the third was any relative with an early onset cancer (FH3), diagnosed before the age of 50. The following chart shows the results of this analysis.

Table 4.3 Results of family history analysis

		Number	FH1 Males p=0.0008 Females p=0.0254	FH2 Males p=0.0113 Females p=0.0005	FH3 males p=0.3051 females p=0.0302
Males	Cases	501	350 (69.9)	103 (20.6)	94 (18.8)
	Controls	491	292 (59.5)	70 (14.3)	79 (16.1)
	Total	992	642 (64.7)	173 (17.4)	173 (17.4)
Females	Cases	452	305 (67.5)	89 (19.7)	115 (25.4)
	Controls	467	281 (60.2)	52 (11.1)	90 (19.3)
	Total	919	586 (63.8)	141 (15.3)	205 (22.3)

From this analysis seems that the second family history definition provides the most significant information about lung cancer risk for both males and females in the MD Anderson case-control study. Although, this information is not incorporated into the smoking based model it shows that a family history component should be added in the future.

4.2 Sources of Age-specific Mortality

The goal of this project is to use a combination of MD Anderson case-control data and external mortality data to fit a carcinogenesis model and estimate the effects of different risk factors on lung cancer development. In order to adjust for the fact that the MD Anderson cases and controls are matched by both age (within 5 years) and smoking status (current, former, and never smokers), data on age-specific mortality by smoking status are needed to adjust for the biases introduced by matching. There are several prospective cohort studies that can provide this information including the American Cancer Society's Cancer Prevention Study I (CPS-I), the Health Professionals Follow-up Study (HPFS) and the Nurses Health Study (NHS). First, the CPS-I study includes tabulated mortality data by race and gender. An examination of these studies, as well as, the age-specific incidence estimates by smoking status will provide clues about which dataset is most reflective of the MD Anderson study participants.

4.2.1 *Cancer Prevention Study 1*

The Cancer Prevention Study I (CPS-I) is a prospective cohort study conducted by the American Cancer Society. There were over 1 million individuals enrolled between

1959 and 1960. Enrollment required participants to be over age 30 and have at least one family member over the age of 45. Study participants completed a baseline survey at enrollment and follow-up questionnaires in years 1961, 1963, 1965, and 1972 allowing for 12 years of follow-up. The baseline questionnaire asked information about health status such as height, weight, demographics, personal and family history of cancer, occupation, diet, alcohol and tobacco use, and physical activity. Follow-up surveys addressed changes in smoking and vital status. The CPS-I study contains information on 456,491 males with known death rates for 117,199, and 594,551 females with known death rates for 88,353 (Thun et al in Smoking and Tobacco Monograph 8). Mortality for 5-year age groups stratified by race, gender, and smoking status for this study is available in Appendix C of Chapter 3 of the Smoking and Tobacco Control Monograph 8.

4.2.2 Nurses Health Study and Health Professionals Follow-up Study

The Nurses Health Study (NHS) is a cohort study on females that began in 1976. Married registered nurses between the ages of 30 and 55 were asked to enroll. The cohort consists of 121,700 female nurses. Every 2 years the participants respond to questionnaires on a wide variety of risk factors of disease including, smoking status, hormone use, and diet, as well as about any health conditions with which they have been diagnosed. Less than 10% of participants have been lost to follow-up.

The Health Professionals' Follow-up study (HPFS) is a cohort study on males that began in 1986. The cohort consists of 51,529 men employed in health professions aged 40-75. As in NHS, the participants fill out questionnaires every 2 years about diseases

and health-related factors such as smoking, and physical activity. Less than 7% of participants have been lost to follow-up.

4.2.3 Comparisons of Mortality Datasets

The table below describes some characteristics of datasets that provide table mortality/incidence rate data stratified by age, gender, ethnicity, and smoking status.

Table 4.4 Cohort characteristics of mortality/incidence rate data

Study	Total	Smoking Status		
		Never	Former	Current
CPS-I males				
Subjects	134,532	92,307	42,225	174,997
Lung Cancer Cases	4,202	215	331	3,656
Avg. follow-up	10.28			
	years			
CPS-I females				
Subjects	398,459	375,649	22,810	158,727
Lung Cancer Cases	602	573	29	621
Avg. follow-up	10.82			
	years			
HPFS				
Subjects	46,050	22,431	19,632	3,987
Lung Cancer Cases	461	58	247	156
Avg. follow-up	14.93			
	years			
NHS				
Subjects	104,493	51,121	24,474	28,898
Lung Cancer Cases	1,165	130	134	901
Avg. follow-up	23.15			
	years			

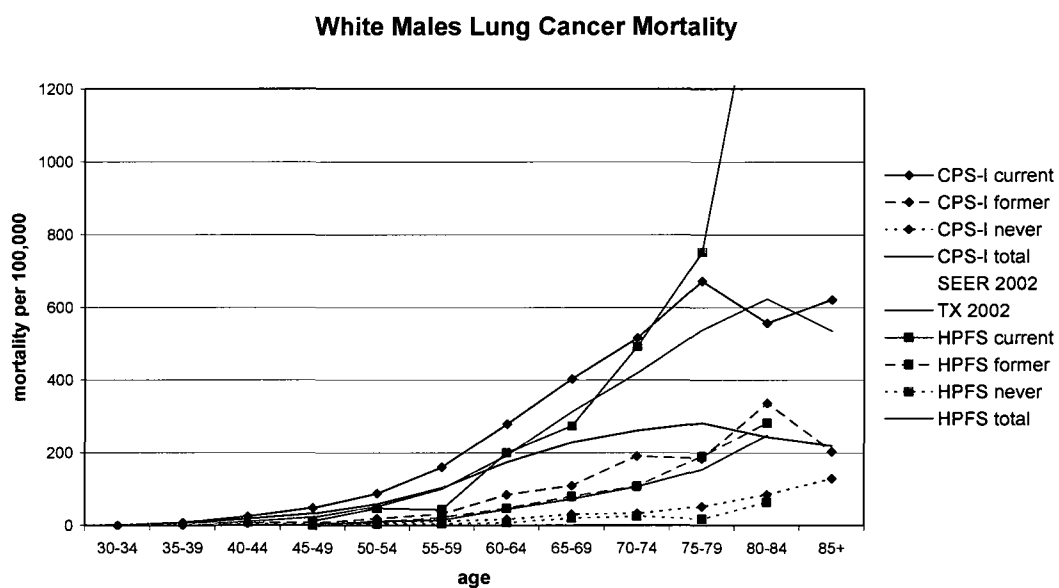


Figure 4.1 Comparison of age-specific mortality rates for white males

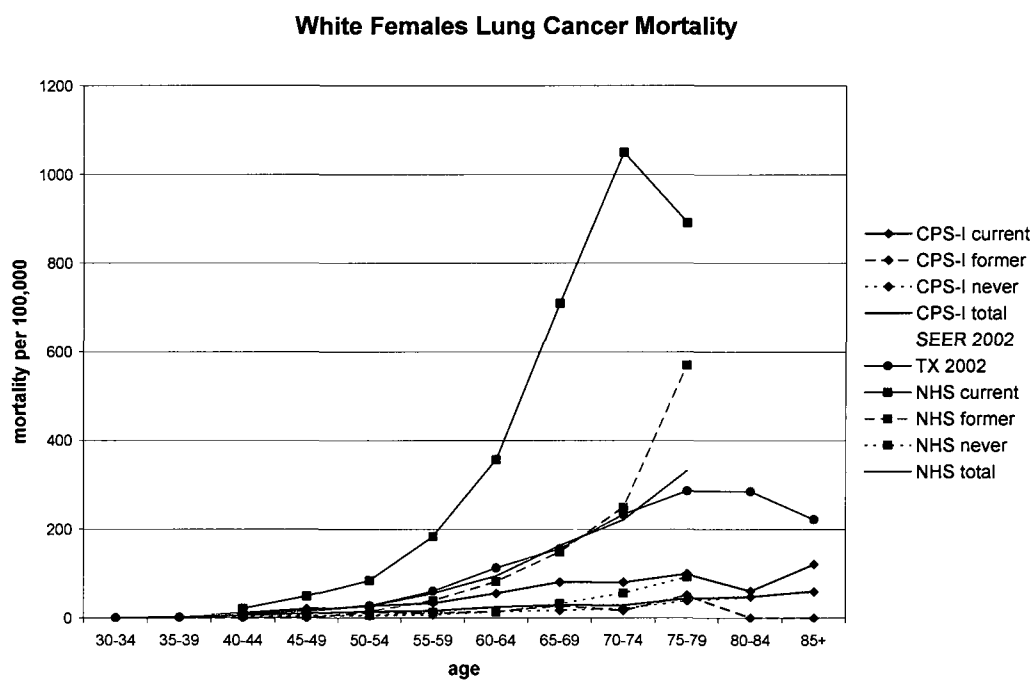


Figure 4.2 Comparison of age-specific mortality rates for white females

The Texas and SEER 2002 mortality rates for lung cancer are similar for both sexes. The SEER 2002 rate is similar to the overall male rate for 1969 through 2004 while for females the 2002 rate is much higher. The increase in female rates is due mainly to the increases in smoking among women and is reflective of the large increase in incidence for females over that time period.

CPS-I rates are consistently lower in females, with the rate for current smokers in females comparable to the overall rate for Texas and SEER in 2002. However, the rates for males are consistent with HPFS, TX and SEER. The overall rate in the HPFS is much lower than both SEER and Texas but this could be accounted for by the lower rates of smoking among the HPFS participants. The never and former smoking rates for both CPS-I and HPFS in males indicates that both studies have comparable rates. However, CPS-I has a higher sample size for smokers indicating that it may be better suited for fitting the model to the data on males.

For females all CPS-I rates are lower than the SEER and Texas rates. The overall rate for CPS-I females is similar to that of the never smokers in that study. This indicates that the CPS-I study rates may not be reflective of current female rates of lung cancer mortality most likely due to the changes in females smoking that have occurred. The overall rates in the Nurses Health Study seems to be similar to the overall SEER and Texas rates indicating that the NHS would be better for fitting the model to the data on females.

4.2.4 *National Health Interview Survey*

The MD Anderson cases and controls are matched by gender, race, age (within 5 years), and smoking status. CPS-I and NHS provide age-specific incidence/mortality rates stratified by gender, race, and smoking status and provide the needed information to adjust for the matching. However, we still require one more piece of information to recreate populational data, namely the proportion of current, former, and never smokers in the population by race and gender.

The National Health Interview Survey (NHIS) is a survey study run by the US Department of Health and Human Services. For this study, annual surveys are conducted in 35,000 to 40,000 households including 75,000 to 100,000 individuals. The survey asks participants about a range of topics including smoking, diet and nutrition, and many other health and wellness related topics. More information on this study can be found at www.cdc.gov/nchs/nhis.htm. This study provides data on the proportion of individuals of each smoking status (current, former, and never) in the population in the year 2000 stratified by gender and race.

4.3 TSCE Model based on Smoking History

Developing a model based on smoking as the only risk exposure allows for comparisons to other smoking based models in the smoking base case project of the CISNET lung group. The MDA case-control data, as well as the LC mortality rates by age and smoking status from CPS-I for males and NHS for females were used to fit the model. A parameterization of 5 fitted parameters with a fixed lag-time is chosen to allow

for the inclusion of other risk variables later and minimize the effects of non-identifiability. The biological parameters of the TSCE model are used for the ease of biological inferences based on the fit. Since the TSCE model is insensitive to choice of lag-time, i.e. the parameters shift in response to changes in lag-time assumptions, a fixed lag-time of 6 years between birth of first malignant cell and death from lung cancer is assumed and is similar to assumed lag-times in other studies (Hazelton 2005) and is in accordance with disease progression models (Fleehinger 1987). The following are the response functions that relate the parameters of the TSCE model depending on smoking intensity, ppd , measured in packs per day.

$$\begin{aligned}
 X &= 10^7 \\
 \nu(t) &= \nu_0 X (1 + a_1 \times \sqrt{ppd}) \\
 \mu(t) &= \nu_0 (1 + a_1 \times \sqrt{ppd}) \\
 \alpha(t) &= \alpha_0 (1 + a_2 \times \sqrt{ppd}) \\
 \gamma(t) &= \alpha(t) - \beta(t) - \mu(t) = \gamma_0 (1 + a_2 \times \sqrt{ppd})
 \end{aligned}$$

The resampling routine needed to be modified because the further matching of the MD Anderson study including smoking status, as well as, age (within 5 years), gender, and race. The routine was modified to include the sampling of smoking status based on rates obtained from the National Health Interview Survey (NHIS) for the year 2000. The following routine outlines the resampling method applied to the MD Anderson case-control data, and incidence/mortality rate data from CPS-I and NHS.

1. The smoking status (Current, Former, Never) is sampled based on rates for the year 2000 from the NHIS. For males: (27.56%, 30.41%, 42.03%) and for females (22.81%, 21.65%, 55.54%) respectively.

2. For each individual, the age bin (5year) in which the person belongs is sampled based on the number of individuals in each age bin of the case-control data for the previously sampled smoking status.
3. Based on the age bin and smoking status generated above, it is randomly sampled whether the individual gets cancer or not based on the estimated probability of an individual within the sampled age bin to get cancer within the 5 years spanning the age bin. This estimate is based on the tabled smoking-status-specific incidence data.
4. Once the age bin, smoking status and cancer status are determined, an individual is sampled from the case-control dataset with the same characteristics and information is used concerning his/her risk factor exposures.

The censoring or age at onset (operationally, the age at LC diagnosis) of the individual is assigned as their age at enrollment from the case-control dataset and the age at entry is assigned as 5 years prior for controls and at a randomly (uniformly) distributed age in the previous 5 years for cases. The resulting fit for a fixed lag-time of 6 years is shown in the following table.

Table 4.5 Parameter fit for the TSCE model based on smoking

Parameter	α_0	γ_0	$\nu_0 X$	a_1	a_2
Males (CPS-I)	2.99 ($\sim 10^{-4}$, 12.54)	0.069(0.064, 0.074)	2.17(1.88, 2.50)	2.66(0.78, 4.86)	0.35(0.08, 0.66)
Females (NHS)	4.60($\sim 10^{-4}$, 20.65)	0.071(0.065, 0.080)	1.93(1.56, 2.19)	2.30(0.36, 4.56)	0.35(0.02, 0.72)

The parameter estimates for males and females show no statistically significant differences. As expected the smoking related variables are statistically significantly positive, showing that smoking does in fact speed carcinogenesis and increase risk for lung cancer. The parameter estimates indicate that smoking one pack-per-day more than triples the mutation rates, and increases the net proliferation rate of initiated cells by 35%.

The following graphs show predicted incidence curves for given smoking histories, lifetime smokers starting at age 18, former smokers starting at age 18 and quitting at age 40, and never smokers. The smoking intensity is assumed to be 1 pack per day.

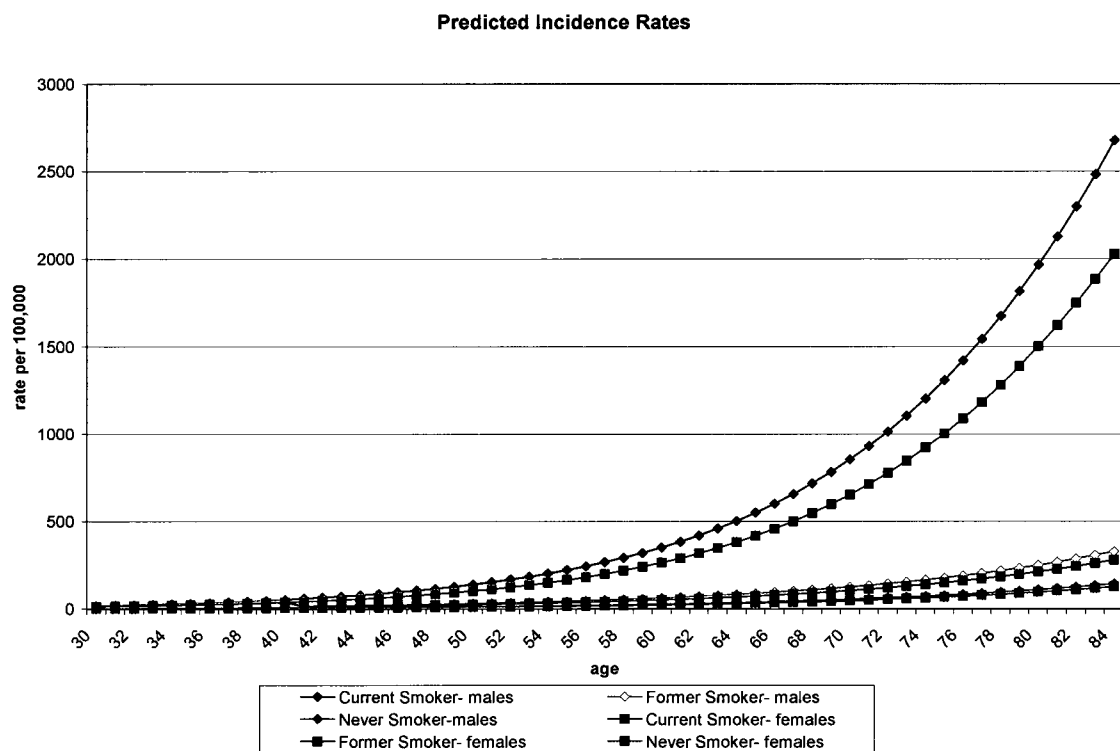


Figure 4.3 Model predicted incidence rates

The predicted incidence rates for former and never smokers are similar in both men and women. However, for life long smokers, males have a slightly higher predicted incidence. Since the parameters don't differ significantly by gender, the predicted incidence can not be shown to differ, which is in agreement with a study of the NHS and HPFS cohorts that showed no statistically significant difference in lung cancer risk in men and women with comparable smoking histories (Bain et al. 2004). However, there was a study based on a CT screened cohort that showed that women have a higher incidence of lung cancer and also higher lung cancer survival (IELCAP investigators 2006). Women are known to have a disproportionately higher percentage of never smokers with lung cancer when compared to men (Subramanian and Govindan 2007) but this may be due to exposures to secondary risk factors.

4.4 Predicting LC Mortality in the CARET study

The estimates were validated by applying the resulting model to data on smoking histories from the non-asbestos exposed control arm of the CARET trial for validation purposes. The smoking based model was used to predict cumulative LC deaths in the non-asbestos exposed control arm of the CARET (Carotene and Retinol Efficacy Trial) study for comparison against the observed LC deaths. The CARET study was a double blind, placebo control trial on the effect of beta-carotene and retinol in the prevention of lung cancer. The non-asbestos exposed group included 7965 men and 6289 women with at least a 20 pack-year smoking history, were aged 50-69 and were current smokers or had quit within the previous 6 years. The study was stopped early when it showed that use of the supplement provided no reduction in cancer. Prior to randomization all

participants were given placebos for 3 months to determine their adherence to taking the vitamins. For this study, data was obtained on the non-asbestos exposed placebo-control arm of the CARET study including data on 6877 individuals (3797 males and 3080 females).

Expected Lung Cancer mortality is calculated based on the individual-level data on smoking history, d_k , and age at enrollment. For each year of follow-up, j , the probability that individual, k , will die from lung cancer is calculated using the following formula where $l_{k,j}$ denotes the length of time that individual k was followed up through year j and $t_{k,j}$ is that individual's age at the start of follow-up year j . The probabilities are conditioned on the fact that participants have not died of lung cancer by age at enrollment, $t_{k,1}$.

$$p_{k,j} = \frac{S(t_{k,j} + l_{k,j}; d_k) - S(t_{k,j}; d_k)}{S(t_{k,1}; d_k)}$$

If an individual was not followed up during the follow-up year j , then $p_{k,j} = 0$.

Then total expected lung cancer deaths, E , in follow-up year j is $E_j = \sum_k p_{k,j}$ and the

cumulative expected deaths, D , by follow-up year j is then $D_j = \sum_{n=1}^j E_n$.

The calculated probabilities $p_{k,j}$, determined the simulation of lung cancer mortality in the CARET study based on the model. The simulation was conducted using the probability of never dying of lung cancer during the study for each individual, $1 - \sum_j p_{k,j}$, and the individual follow-up year probabilities of lung cancer death, $p_{k,j}$.

In a single simulated study, whether or not lung cancer death occurs is simulated for each individual. If an individual does die from lung cancer then the year of follow-up at which the individual dies is simulated according to the probabilities. The cumulative number of lung cancer deaths per follow-up year is calculated for each simulated study. The simulation is repeated 5,000 times to compare expected lung cancer mortality and produce confidence intervals.

The mean predicted cumulative LC mortality as well as confidence limits are depicted in the following figures. The model produces accurate predictions for overall lung cancer mortality in the CARET study. There were 364 LC deaths in the course of the CARET study and the model predicted 366.8 (CI 331-402). The model was also able to predict LC deaths for males and females within confidence limits as shown in the following chart.

Table 4.6 CARET study predicted and observed LC deaths

	Overall	Males	Females
Observed	364	225	139
Predicted	366.8(331,402)	236.8(208,266)	130.0(109,152)

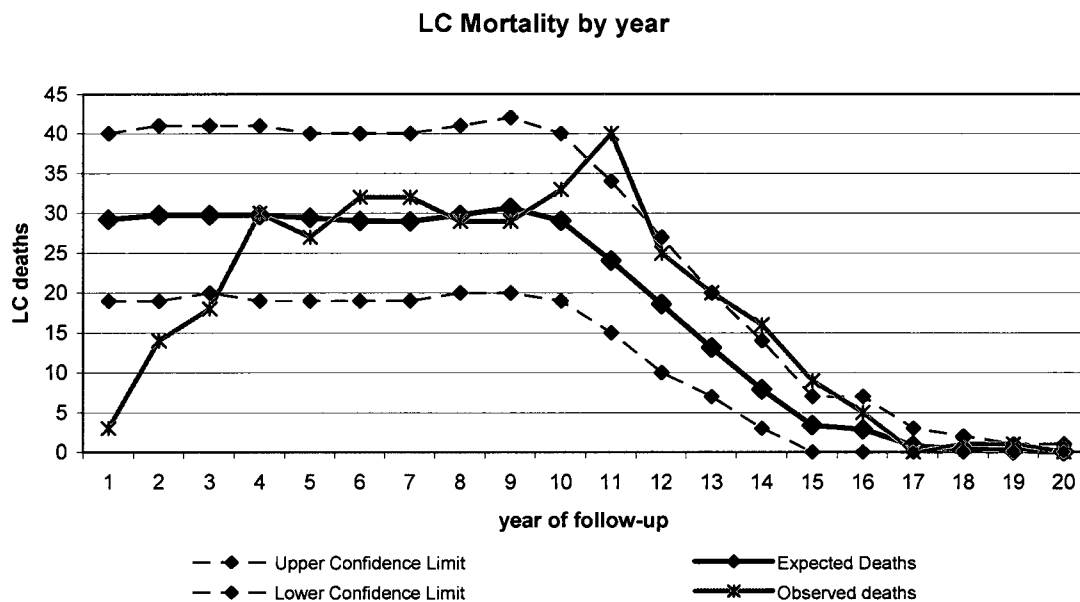


Figure 4.4 Yearly predicted and observed LC deaths- CARET

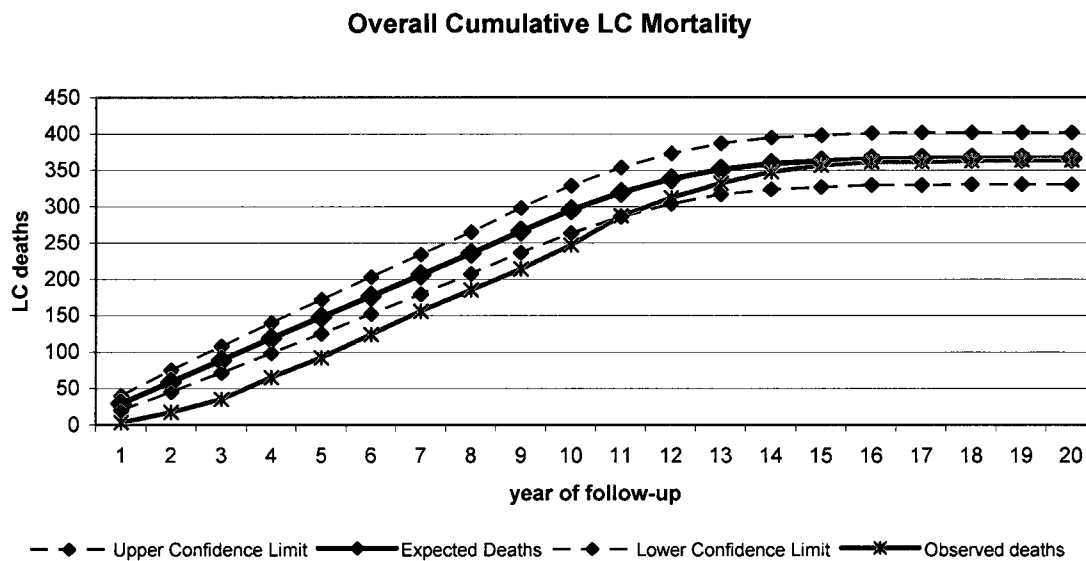


Figure 4.5 Cumulative predicted and observed LC deaths- CARET

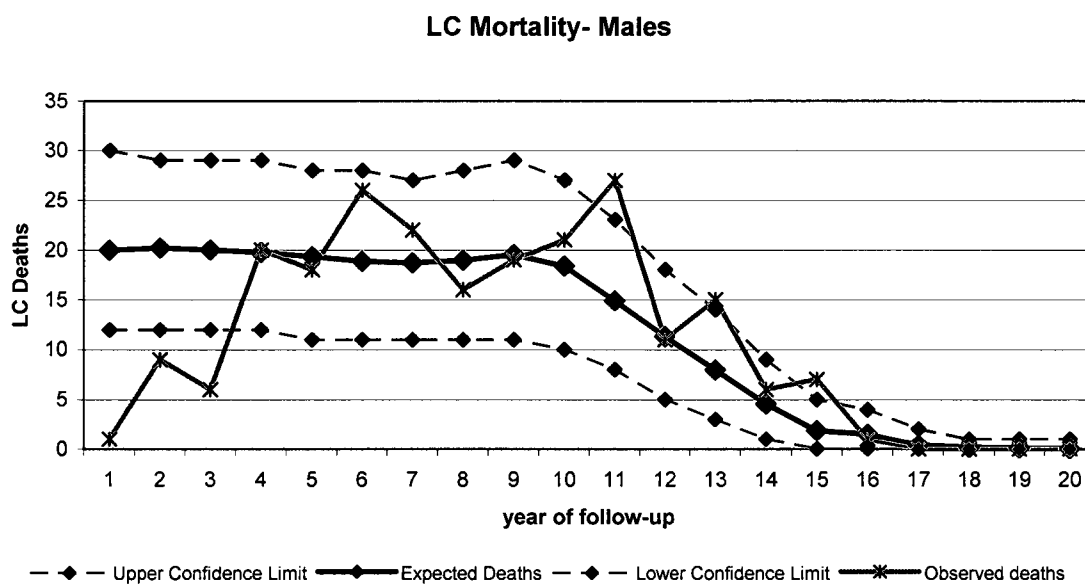


Figure 4.6 Yearly predicted and observed LC deaths- CARET males

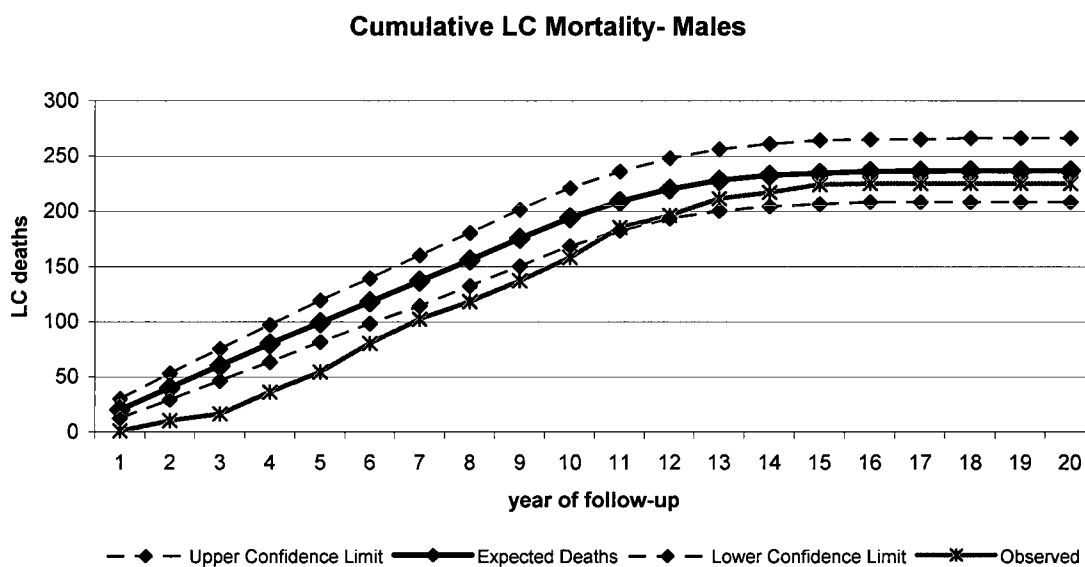


Figure 4.7 Cumulative predicted and observed LC deaths- CARET males

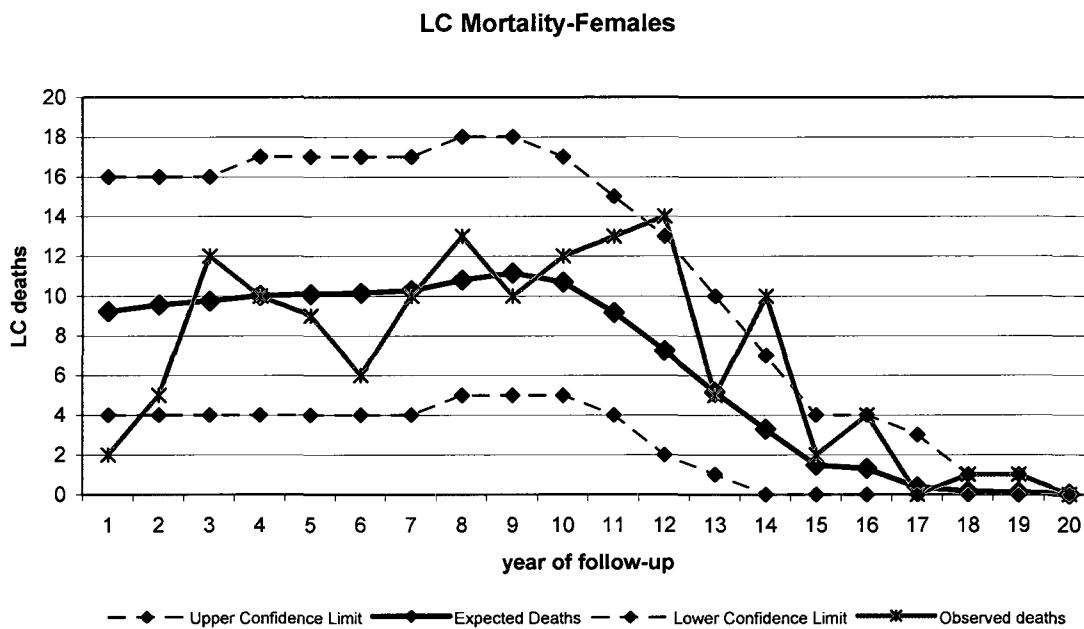


Figure 4.8 Yearly predicted and observed LC deaths- CARET females

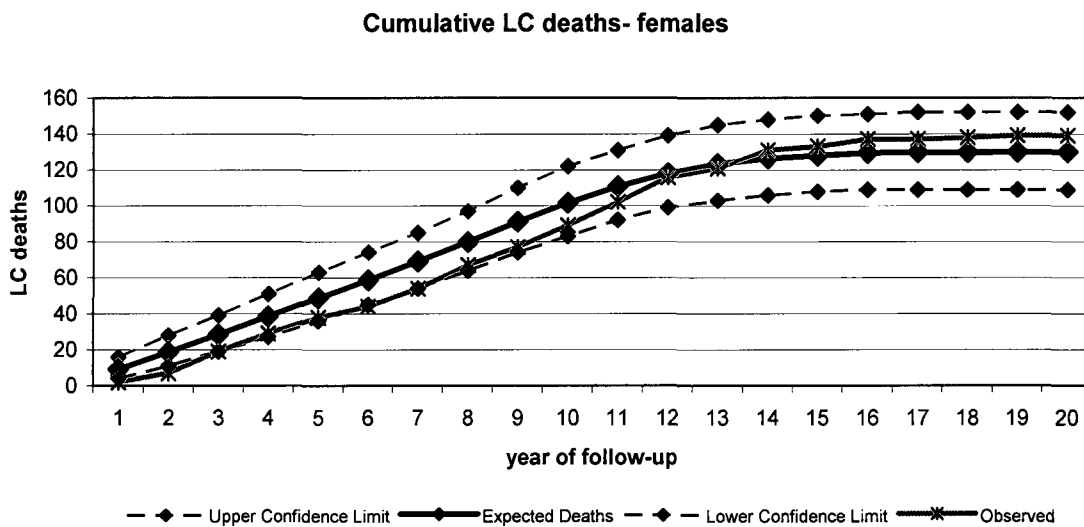


Figure 4.9 Cumulative predicted and observed LC deaths- CARET females

The yearly number of LC deaths predicted and observed suggest that there may be some healthy volunteer bias in the participants of CARET. Observed LC mortality was lower than predicted for the first year in females and in years 1-3. However, at about year 10 of follow-up we start to see the opposite trend. After about year 10, there are some years with more observed LC mortality than predicted. This effect could be seen when healthier participants are lost to follow-up more often than unhealthy participants. Less healthy who know that they are at high risk may be more inclined to remain in the study than participants who feel healthy. Those who feel healthy later may not see the benefit of remaining in the study and thus be more likely to drop out. These effects seem to balance each other by the completion of follow-up.

Removing these effects to determine the models ability to predict LC deaths is important. In order to fully remove healthy volunteer effect the first 3 years are removed. The model is used to predict LC mortality in years 4-10 for comparison against the observed. This time interval is used because it removes any possible healthy volunteer effect as well as any underestimation in the later years. It is also the same interval that will be used in the CT screening simulation. The following graphs show the observed and predicted LC deaths in years 4-10. The model seems to do very well in predicting lung cancer mortality in these years.

Table 4.7 CARET predicted and observed LC deaths (years 4-10)

	Overall	Males	Females
Observed	212	142	70
Predicted	207.0(180,235)	133.6(112,156)	73.4(57,91)

Cumulative LC Deaths Adjusted for Healthy Cohort Bias (years 4-10)

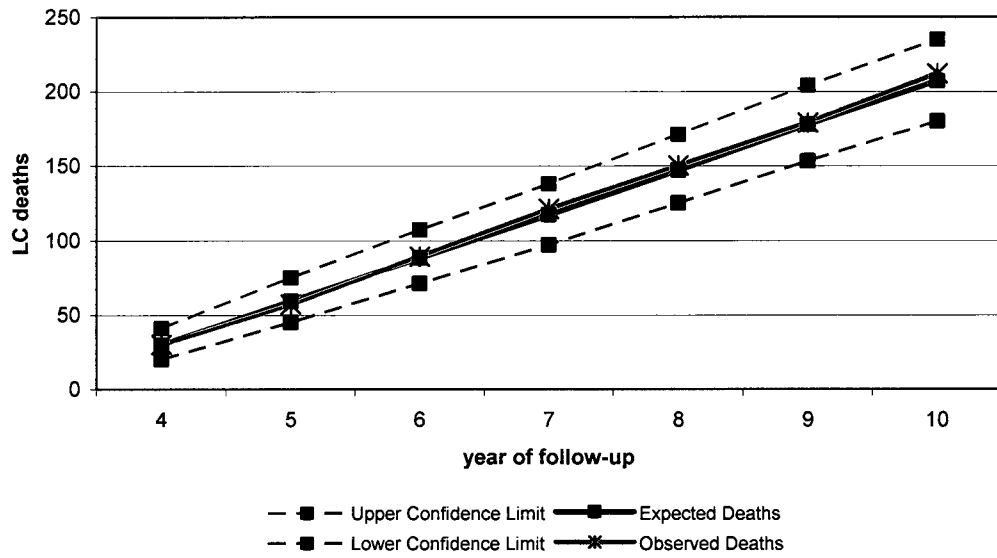


Figure 4.10 Predicted and observed cumulative LC deaths- CARET (years 4-10)

Cumulative LC Mortality adjusted- Males (years 4-10)

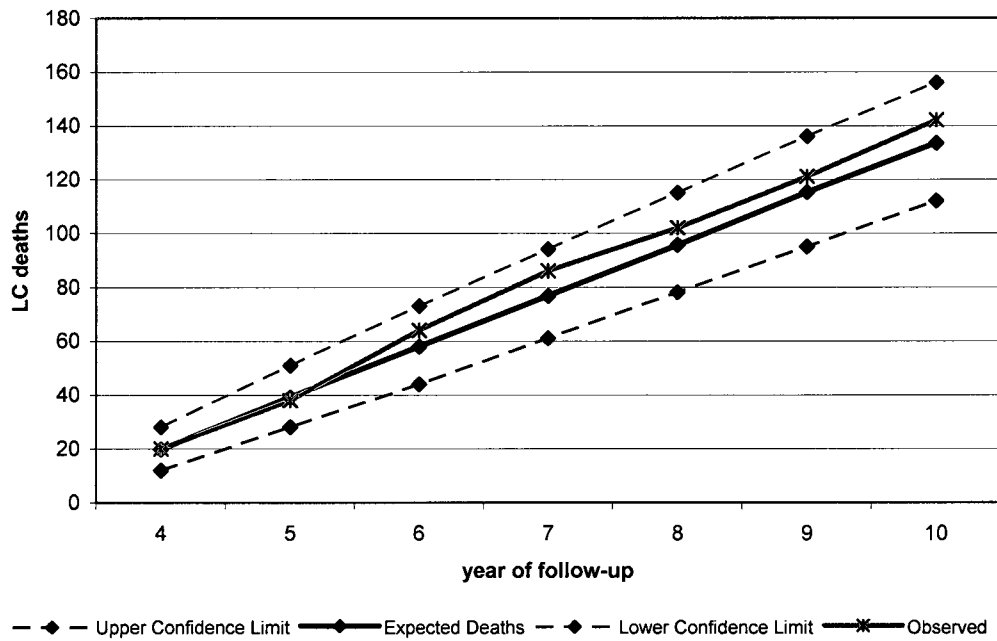


Figure 4.11 Predicted and observed cumulative LC deaths- CARET males (years 4-10)

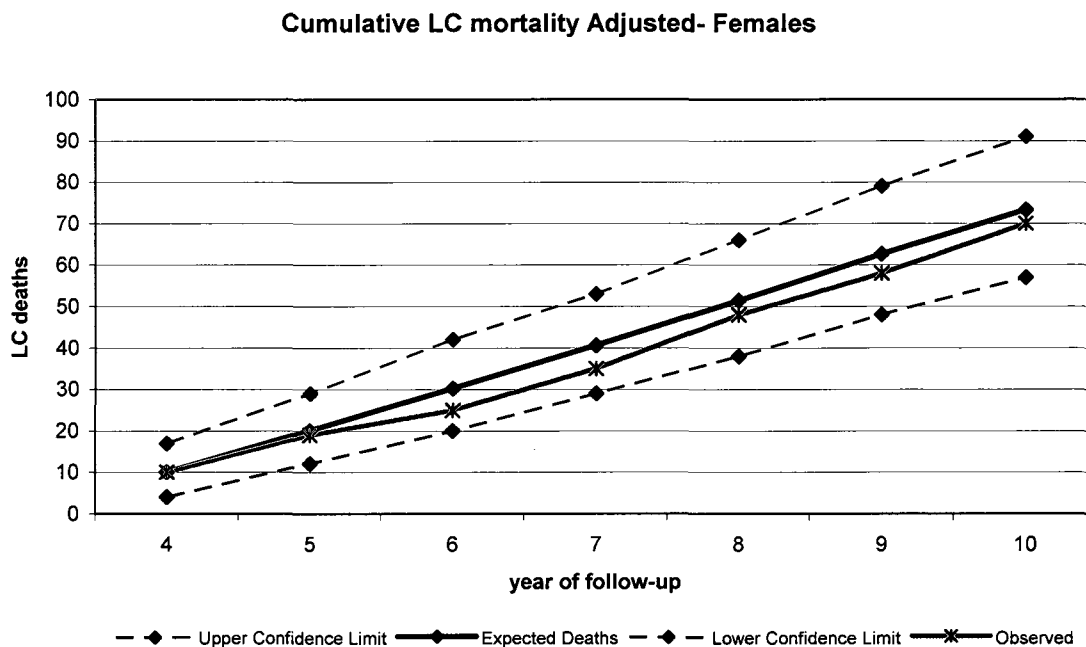


Figure 4.12 Predicted and observed cumulative LC deaths- CARET females (years 4-10)

4.5 Simulating US LC mortality

The smoking history generator provided by CISNET and the model based on smoking, are used to simulate LC mortality in the US population for comparison against the observed. Given year of birth, gender and race, the smoking history generator provides a smoking history and age of death t_d from causes other than lung cancer. Using each individual's unique smoking history and death of other cause times, LC mortality is simulated according to the model based on the parameter estimates obtained by the resampling method, using the following simulation routine suggested by Kaiser and Heidenreich (2004).

1. Then the probability of not developing cancer by the age at death from any other cause t_d was calculated according to the TSCE model: $p_{TSCE} = S(t_d)$.
2. Then a uniform(0,1) random variable, u , was drawn.
 - a. If $u \leq p_{TSCE}$ then time of censoring is t_d and no cancer develops during the individual's lifetime.
 - b. If $u > p_{TSCE}$ then cancer develops during the individual's lifetime and is diagnosed at age, t , computed by inverting the survival function, $u = S(t)$.

50,000 individuals were simulated per birth cohort 1891-1970. After the birth cohorts are combined the year-by-year age distributions are adjusted using re-weighting to match the observed US population. LC mortality for males and females were simulated separately. The following graphs show the observed and simulated LC mortality for years 1985-2000.

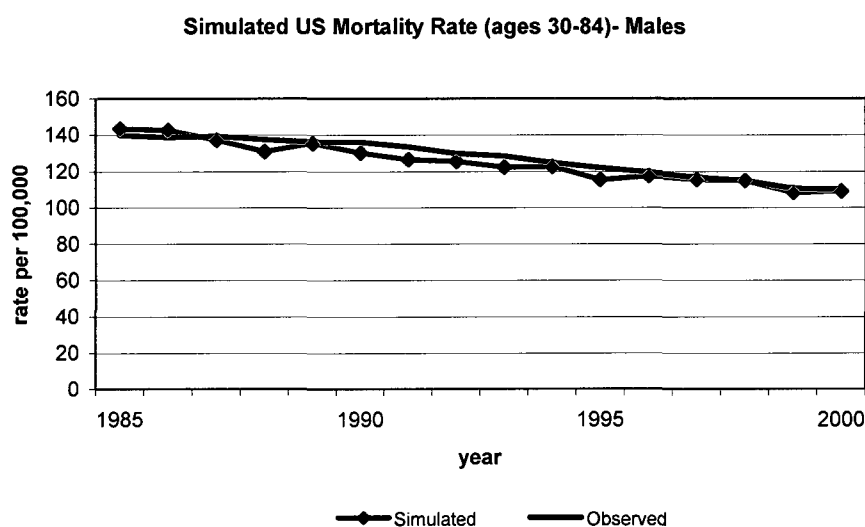


Figure 4.13 Simulated and observed US LC mortality rates- males

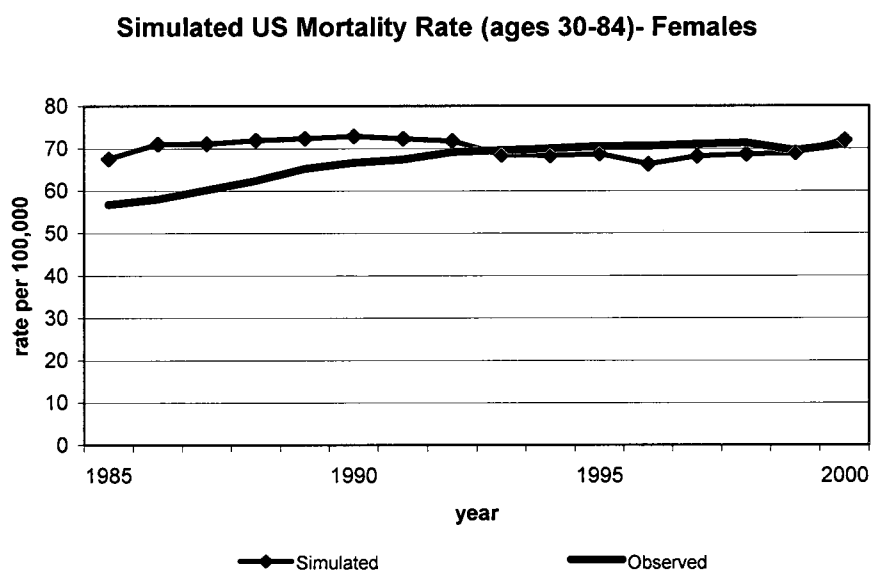


Figure 4.14 Simulated and observed US LC mortality rates- females

The results from this simulation show that the model can reasonably re-create US LC mortality. Simulation of LC mortality in the US population shows an acceptable prediction of US LC mortality in the period 1985-2000. However, if the mortality is simulated for years 1975-1985 we find that the model predicts higher than observed LC mortality. This could be reflective of the changes in cigarettes and smoking behaviors that occurred in the period 1950 through the mid 1980s, while the model was fit to more contemporary data. For women, the higher predictions persist through the year 1992 after which the predictions are accurate.

Starting in the 1950s changes were made to cigarettes and tobacco that included the introduction of filters, changes in the method of curing tobacco, and marketing of low tar cigarettes with reduced nicotine levels. Reduced tar and nicotine cigarettes sought to lower the carcinogen exposure produced by cigarette smoking. While the changes did reduce levels of tar and some carcinogens other carcinogens were increased (Hoffman

1997). Also, there were observed smoking behavior changes in smokers of lower yield cigarettes including smoking more cigarettes per day, taking more puffs from each cigarette, and inhaling more deeply (Thun 2001) referred to as compensatory smoking. Compensatory smoking is driven by the need to maintain nicotine levels to satisfy the nicotine addiction. By smoking more and inhaling further into the lungs, nicotine levels are increased and exposures to the carcinogens in cigarettes are higher. In the mid 1960s a new method of curing tobacco was employed that lead to a marked increase in tobacco-specific N-Nitrosamines. These changes in cigarettes that started in 1950 and continued through the mid 1980s increased the carcinogenicity of modern day cigarettes. Thus, one of the limitations of our model is that it does not capture the effect of changes in cigarettes and smoking behaviors over time.

4.6 CISNET Smoking Base Case Project

The CISNET lung group's smoking base case projects uses simulation to estimate the number of LC deaths that were averted by tobacco control policies that were initiated after the surgeon general's report of 1964 warning on the dangers of tobacco use. The project also simulates the deaths that would have been averted if smoking had been banned.

The smoking history generator can be used to simulate population smoking trends based on 3 different scenarios. The first is based on the NHIS, referred to as the Actual scenario, and re-creates observed trends in smoking and mortality. The second is based on what would the expected smoking would be if there were no public health information about the dangers of smoking. It attempts to simulate smoking histories as if there were

no known information about the dangers of smoking and individuals continued smoking along the patterns observed before the 1964 surgeon general's report (US Department of Health, Education, and Welfare 1964). The last scenario is based on the premise that smoking is banned in 1965 after the surgeon general's report came out. For this scenario all individuals quit smoking in 1965 and there is no smoking from that point on. The goal of having the 3 scenarios is to quantify the effect of the tobacco control programs that were initiated after the surgeon general's report as compared to what would have happened if individuals were forced to quit.

The following graphs show the simulated LC mortality under these 3 scenarios based on the smoking history generator.

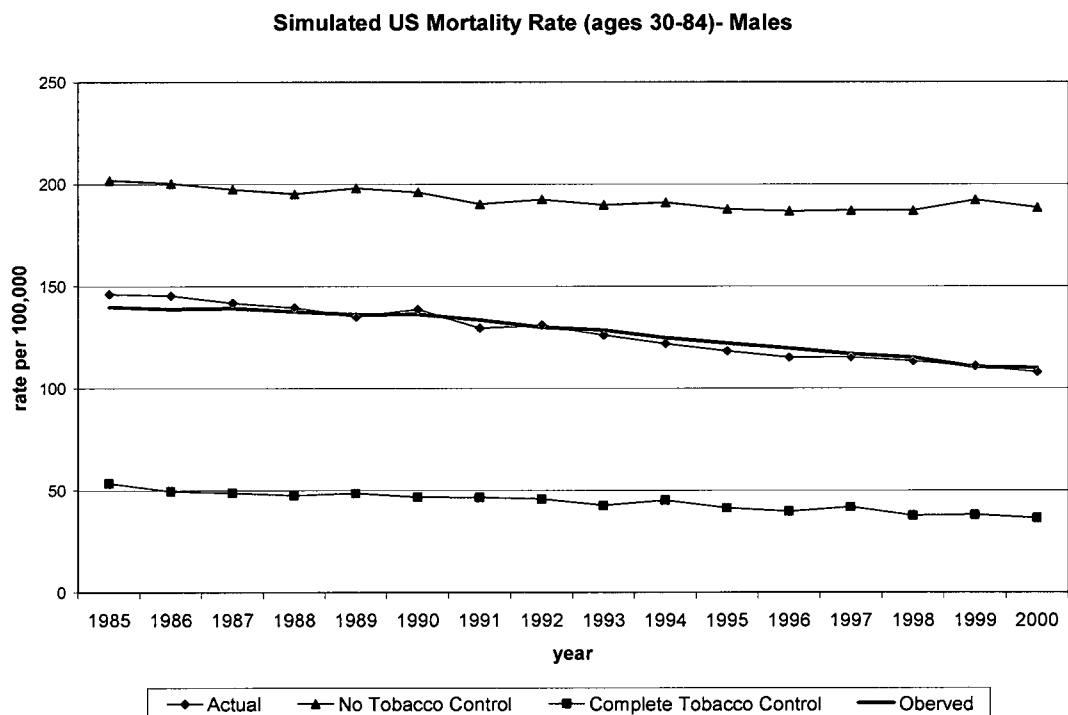


Figure 4.15 CISNET Smoking Base Case Simulation- males

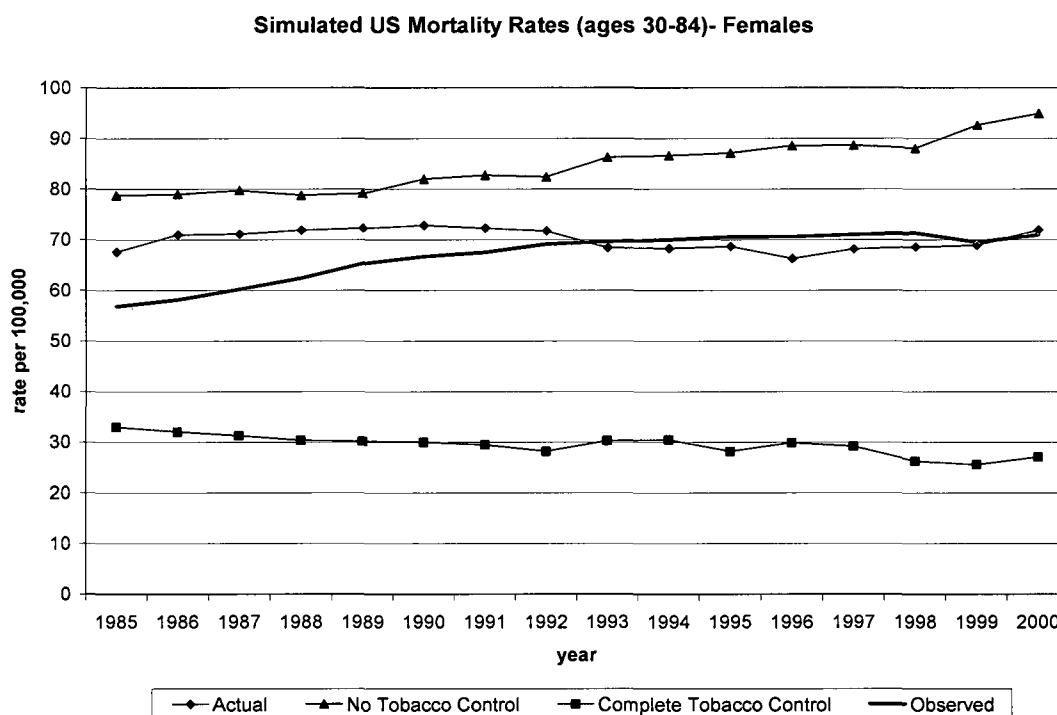


Figure 4.16 CISNET Smoking Base Case Simulation- females

The difference between the No Tobacco Control and Complete Tobacco Control lines provides a measure of the potential number of LC deaths that could have been avoided if all smokers were forced to quit. The difference between the No Tobacco Control and simulated Actual LC deaths provides a measure of the realized reduction in LC deaths accomplished through tobacco control policy. As seen from the graph below, the tobacco control policies only met 40% of the total potential lung cancer death reduction in the years 1985-2000.

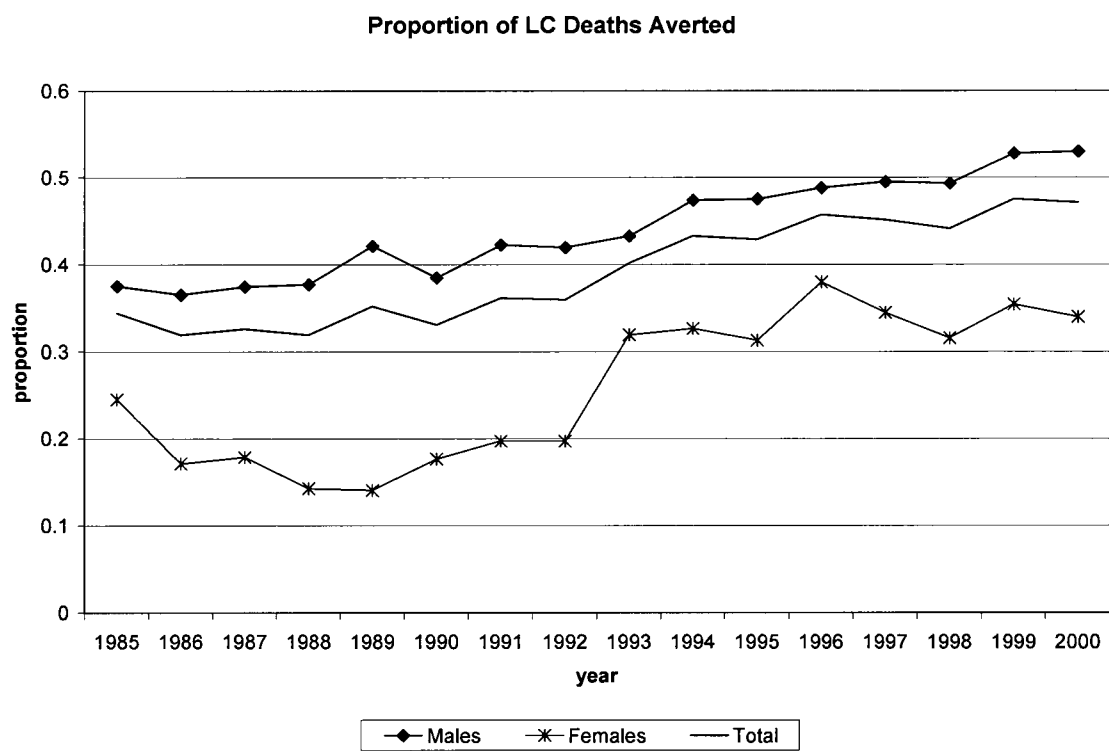


Figure 4.17 Proportion of LC deaths averted through tobacco control policies

Chapter 5

Implications for Screening

5.1 Single-arm CT Screening Trial for LC

The model is used to predict LC mortality in the absence of screening to use as a surrogate control arm for comparison against observed LC mortality in a single-arm trial on CT screening. The main criticism of the ongoing ELCAP CT screening study is that it lacks a control arm. The model is used to simulate the expected LC mortality in the absence of the CT screening intervention. Using data on gender, age, and detailed smoking history from participants in the ELCAP study, LC mortality is simulated for the individuals as done in the CARET study analysis.

5.2 NY LC Screening Study

The New York study uses CT scans as a screening tool for Lung Cancer. Study participants are aged 55 or older, have at least a 10 pack-year smoking history, and have no symptoms of LC at time of enrollment. Detailed smoking histories including age at initiation, age at cessation, and number of cigarettes per day smoked, were obtained from questionnaires at the time of enrollment. The final data set includes 7994 individuals.



Figure 5.1 Person-years in the NY cohort per year of follow-up

5.3 Simulating Mortality in the Absence of Screening

The model used predicts the risk of an individual dying from lung cancer in a given time interval based on individual's unique smoking history and age. The probability that an individual will not die of lung cancer by age t , is defined as the survival probability and the function as $S(t)$. In this model $S(t)$ depends on the individuals smoking history with age at initiation of smoking, i , age at cessation of smoking, c , and number of cigarettes smoked per day, s , and will be referred to as $S(t; i, c, s)$.

Expected lung cancer mortality for the study is calculated based on the individual-level data on smoking histories and ages at enrollment provided for the NYC cohort. For each year of follow-up, j , the probability that individual, k , will die from lung cancer is calculated using the following formula where $l_{k,j}$ denotes the length of time that individual k was followed up through year j and $t_{k,j}$ is that individual's age at the start of

follow-up year j . The probabilities are conditioned on the fact that participants have not died of lung cancer by age at enrollment, $t_{k,1}$.

$$p_{k,j} = \frac{S(t_{k,j} + l_{k,j}; i_k, c_k, s_k) - S(t_{k,j}; i_k, c_k, s_k)}{S(t_{k,1}; i_k, c_k, s_k)}$$

For individuals not followed up during the follow-up year j , $p_{k,j} = 0$.

Then total expected lung cancer deaths, E , in follow-up year j is $E_j = \sum_k p_{k,j}$ and the

cumulative expected deaths, D , by follow-up year j is then $D_j = \sum_{n=1}^j E_n$.

The calculated probabilities $p_{k,j}$, allowed for the simulation of expected lung cancer mortality in the NYC cohort based on the model. The simulation was conducted using the probability of k -th individual not dying of lung cancer during the study, $1 - \sum_{j=1}^{10} p_{k,j}$, and the individual's follow-up year probabilities of lung cancer death, $p_{k,j}$, $j=1, \dots, 10$, over the 10 years of total follow-up time.

Using the probabilities defined above, lung cancer mortality is simulated for each individual. If the simulated individual does die from lung cancer then the year of follow-up at which the individual dies is simulated. The cumulative number of lung cancer deaths per follow-up year is then calculated for each simulated study. The simulation is repeated 5,000 times to compare expected lung cancer mortality and produce confidence intervals. The confidence intervals are estimated using the 2.5% and 97.5% of the 5,000 simulated studies.

5.4 Results of Screening Simulation

The following figure provides the yearly number of expected and observed deaths from lung cancer in the NYC cohort together with the 95% confidence interval. The yearly predicted and observed LC deaths are also provided in the following table to show the magnitude of any healthy volunteer bias. It appears that the healthy volunteer effects through year 3, decreasing in magnitude from year 1 to year 3. For the main analysis the first 3 years of follow-up are removed to conservatively adjust for any possible healthy volunteer effect.

Table 5.1 Observed and predicted LC deaths in years 1-4 in the NYC cohort

Year of follow-up	Observed	Predicted	95% CI
1	6	36.3	(25,48)
2	13	36.7	(25,49)
3	16	24.3	(15,35)
4	12	14.5	(7,23)

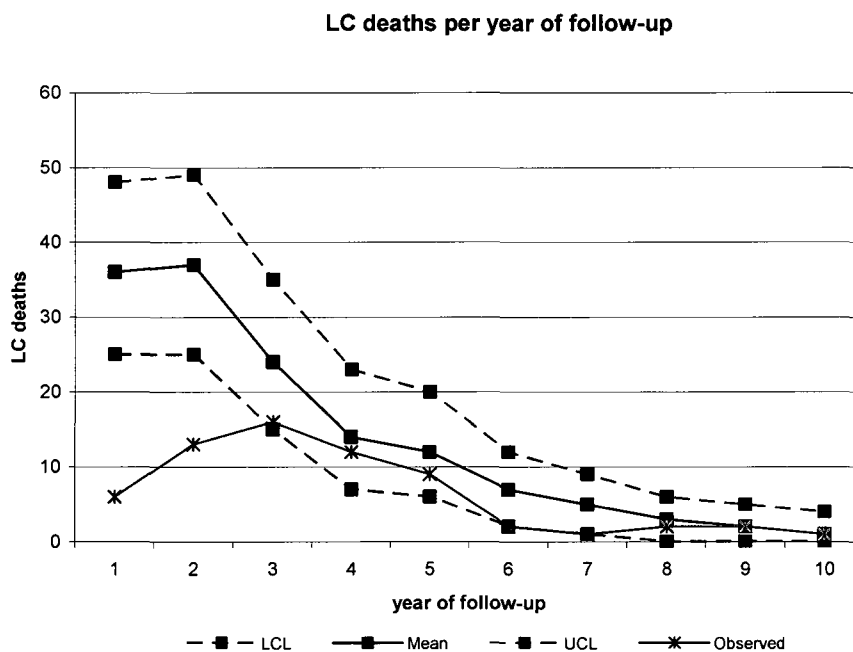


Figure 5.2 Yearly simulated and observed LC deaths in the NYC CT screening cohort

The following graphs show the observed and predicted cumulative LC deaths starting in year 4 and ending at the end of follow-up (year 10).

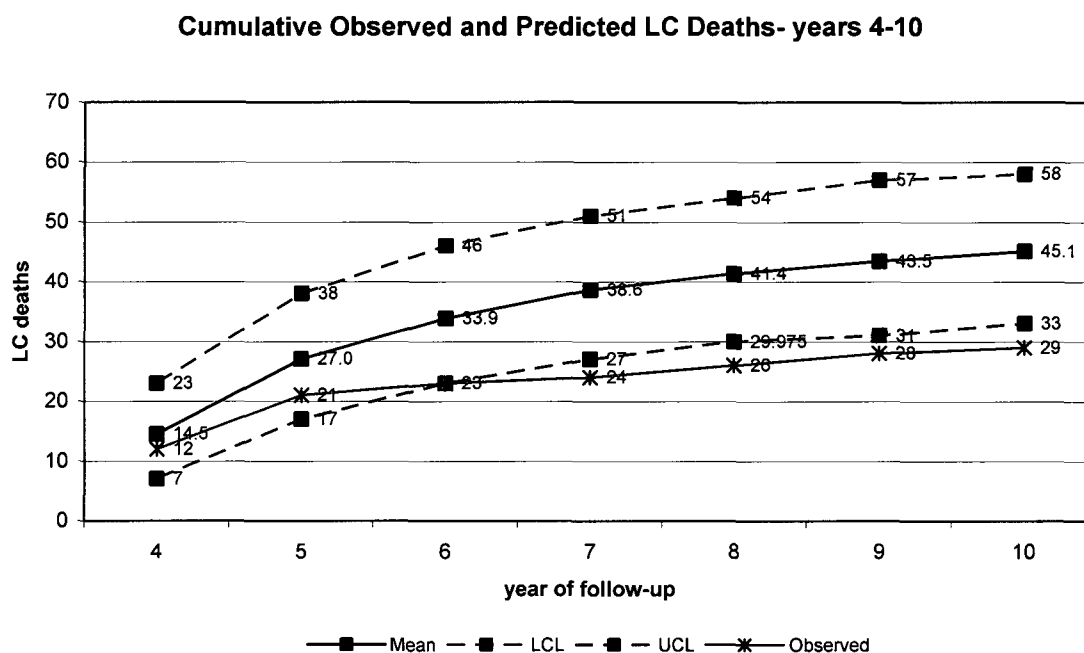


Figure 5.3 Predicted and observed cumulative LC deaths in the NYC cohort (years 4-10)

The expected number of LC deaths for years 4-10 (first 3 years are excluded to adjust for healthy volunteer bias) is 45.1 and the 95% confidence interval ranges from a lower bound of 33 to an upper bound of 58. The observed number of LC deaths for follow-up years 3-10 in the screened cohort was 29.

Another approach to determining the mortality reduction is to calculate the SMR. The SMR was significant for years 4-10 when the deaths of participants who dropped out of screening were included, $29/45.1 = 0.64$, showing a mortality reduction of 36% (95% CI: 0.12, 0.50, $P < 0.001$). If the first 2 years of follow-up are removed, there is still a statistically significant estimated mortality reduction of 35% (95% CI: 0.17, 0.48,

$p < 0.001$), which is very close to the estimated mortality reduction if the first 3 years are excluded, indicating that healthy volunteer effect is fully removed by year 3.

For the NYC cohort, the model predicts 45.1 LC deaths over the years 4 through 10 of follow-up after adjusting for healthy volunteer bias in the first 3 years, compared to the 29 observed deaths in those years. Even conservatively dealing with healthy volunteer effect it appears that CT screening for lung cancer does provide a mortality benefit.

5.5 CT Screening Controversy

There was recently a similar analysis conducted by Bach et al (2007) on 3 different CT screening cohorts including Instituto Tumori, the Mayo Clinic, and the Moffit Center which showed contradictory results to what we found in this study, showing no mortality benefit to CT screening for lung cancer. However, the confidence limits showed there could be as much as a 30% reduction in LC mortality. Some of the reasons for discrepancies between this study and our study could include that this previous study contained mortality data on only 3,210 individuals who had a median follow-up time of 3.7 years as compared to this analysis containing data on 7,995 individuals with a slightly longer median follow-up time of 4.4 years. The Bach study also included more outcomes such as LC diagnosis and surgical resections, while this study focuses on LC mortality alone. The studies also differ in the modeling used to estimate risk of lung cancer death.

The efficacy of CT scanning as a screening tool for lung cancer remains a controversial topic. As more data comes in, it should become apparent whether it can be reliably used to lower LC mortality. If it is proven effective then more analysis will be

needed to determine which individuals should be screened, how often they should be screened, and how should the nodules found be managed. The results from the randomized National Lung Screening Trial are anticipated but may not fully answer questions about efficacy because instead of having standard care (no recommended screening) the control group is being screened with x-ray, which will bias results towards underestimating any observed mortality benefit.

Chapter 6

Discussion and Future Directions

6.1 Summary

In this thesis, a new method of reconstructing time to event data from the combination of case-control data on risk factors and tabled age-specific incidence/mortality rate data. This method is based on the assumption that given the matching stratum and cancer status, cases and controls are randomly sampled from the population. This method is applied to fit a two-stage clonal expansion model for lung carcinogenesis. Simulation studies showed that the method resulted in reasonable fits for this model.

A smoking based TSCE model was fit to MD Anderson case-control data on smoking histories and tabled age-specific mortality rate data stratified by smoking status from CPS-I for males, and NHS for females. The model was fit separately for males and females, and the resulting parameter estimates do not statistically differ by gender.

The model was then validated against the control group of the non-asbestos exposed, also known as the heavy smoker cohort of CARET. The model predicted 366.6 LC deaths over the course of the study compared to the observed 364 resulting in an accurate fit. The model was able to reasonably reproduce US mortality rate trends for the

years 1985-2000 using simulation. The model was also applied to the CISNET Lung Group's smoking base case project.

The model was then used to simulate LC mortality in the absence of screening, for use as a comparison control for a single arm trial of CT screening for lung cancer. The results of this effort show a 36% reduction in mortality resulting from CT screening intervention.

6.2 Future Considerations

The model can be further expanded in a number of ways. Most basically, by including other known risk factors for LC including family history, presence of lung disease such as COPD, exposures to dusts and asbestos, and genetic factors. Also, within the TSCE model framework the lag-time from appearance of first malignant cell to lung cancer diagnosis/death is assumed and the model is fit. The model itself is insensitive to choice of lag-times as the fitted parameters will adjust based on the assumptions. A TSCE model fit based on no lag-time will produce similar incidence and risk predictions to a TSCE model fit with different lag-time assumptions fit to the same data. In this study the lag-time is assumed to be a fixed length of 6 years. This is a reasonable assumption when compared to other studies. However, the lag-time itself could be modeled independently based on what is known about tumor growth and progression rates. Modeling the lag-time separately could also provide a framework for predicting the tumor size, growth rate, and histological sub-type and thus provide a more encompassing modeling framework.

Healthy volunteer bias is a well documented effect (Austin et al. 1981, Benfante et al. 1989, Bisgard et al. 1994, Lindsted et al. 1996, Chou et al. 1997, Etter et al. 1997, Froom et al. 1999, and Hara et al. 2002) in cohort studies enrolling volunteers. Volunteers for the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) tend to be more physically active, have higher education, and are less likely to be overweight than the general population (Pinsky et al. 2007). Also, a study involved in recruiting smokers still showed significant healthy volunteer effect (Thomson et al. 2005). The recruited smokers had on average 15.45 hours of physical activity compared to the national average of less than 1 hour per week. Volunteer of this study also were less likely to be overweight, drank less alcohol, and had lower rates of hypertension and diabetes than the US national average. Every cohort is different with regards to recruitment, but studies involving volunteers are likely to suffer from the healthy volunteer effect. Although, the effect is well documented there is no current method designed for adjusting or removing the effect in analyses. Development of a method that could adjust for the healthy volunteer effect, perhaps by removing proportions of predicted deaths decreasing in magnitude over the first few years for a study, would help improve the predictions of models as well as estimate the effectiveness.

Primary tumors can be located in different parts of the lung. Detection of more than one lesion is relatively common. The number and localization of the tumors is an important clinicopathological feature; depending on the location, a tumor of the same size can produce very different symptoms and outcomes. Factors affecting the spatial distribution of lung lesions are poorly understood. A study of these factors can shed light on lung cancer risk and outcomes. Using information on risk factors and tumor sizes and

locations, it is possible to model the spatial distribution and sizes of the tumors. It is also possible to further study how tumor characteristics (location, size, and number of lesions), influence lung cancer symptoms and prognosis. A complication of this analysis is that the presence of multiple tumors currently affects the stage classification and recommended treatment which in turn affects outcomes. Having knowledge about where a tumor is likely to develop and how fast it is likely to grow, based on a patient's risk factor history would be valuable in developing LC screening programs.

In 1953, Slaughter et al introduced field cancerization as a theory about the spread of pre-malignant "fields" which give rise to tumors in oral cancers. Using data on lung cancer patients including data on epidemiological risk factors as well as tumor sizes and locations it is possible study the evidence for field cancerization in the lung by studying the locations of synchronous primary lung cancers. The field cancerization process has been in more accessible cancer sites (Braakhuis et al. 2003, 2004) including oral cancer (Copper et al. 1993). Genetic studies of the lung are split on whether the field exists in lung cancer (Franklin et al. 1997, Sozzi et al. 1995). Evidence of field cancerization may explain the presence of multiple lesions, and may also provide insight on the mechanism of recurrence for lung tumors.

References

Aberle DR, Brown K: Lung cancer screening with CT. *Clin Chest Med* 29: 1-14, 2008

Alberg AJ, Samet JM: Epidemiology of Lung Cancer. *Chest* 123(1 supplement): 21S-49S, 2003

Alberg AJ, Brock MV, Samet JM: Epidemiology of lung cancer: Looking into the future. *J Clin Oncol* 23: 3175-3185, 2005

Armitage P, Doll R: The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* 8: 1-12, 1954

Armitage P, Doll R: The two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *British Journal of Cancer* 11: 161-169, 1957

Austin MA, Criqui MH, Barrett-Conner E, Holdbrook MJ: The effect of response bias on the odds ratio. *Am J Epidemiology* 114:137-43, 1981

Bach PB, Jett JR, Pastorino U, et al: Computed tomography screening and lung cancer outcomes. *JAMA* 297: 953-961, 2007

Bach PB, Elkin EB, Pastorino U, et al: Benchmarking lung cancer mortality rates in current and former smokers. *Chest* 126: 1742-1749, 2004

Bailey-Wilson JE, Amos CI, Pinney SM, Peterson GM, et al: A major susceptibility locus maps to chromosome 6q23-25. *Am J Hum Genet* 75(3): 460-474, 2004

Bain C, Feskanich D, Speizer FE, Thun M, Hertzmark E, Rosner BA, Colditz GA: Lung cancer rates in men and women with comparable histories of smoking. *J Natl Cancer Inst*, 96: 826-34, 2004

Bartoszynski R, et al: Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis. *Math Biosci* 171: 113-142, 2002

Benfante R, Reed D, MacLean C, Kagan A. Response bias in the Honolulu Heart Program. *Am J Epidemiol* 130:1088-100, 1989

Bisgard KM, Folsom AR, Hong C, Sellers TA. Mortality and cancer rates in nonrespondants to a prospective study of older women: 5-year follow-up. *Am J Epidemiol* 139:990-1, 1994

Black WC, Baron JA: CT screening for lung cancer: Spiraling into confusion? *JAMA* 297(9): 995-997, 2007

Boffetta P, Pershagen G, Jockel KH, et al: Cigar and pipe smoking and lung cancer risk: a multicenter study from Europe. *J Natl Cancer Inst* 91: 679-701, 1999

Boffetta P, Nyberg F: Contribution of environmental factors to cancer risk. *Br Med Bull* 68: 71-94, 2003

Braakhuis BJM, Tabor MP, Kummer JA, Leemans CR, Brakenhoff RH: A genetic explantation of Slaughter's concept of field cancerization: evidence and clinical implications. *Cancer Research* 63: 1727-1730, 2003

Braakhuis BJM, Leemans CR, Brakenhoff RH: A genetic progression model of oral cancer: current evidence and clinical implications. *J Oral Pathol Med* 33:317-22, 2004

Brennan P, Buffler PA, REnolds P, et al: Secondhand smoke exposure in adulthood and risk of lung cancer among never smokers: A pooled analysis of two large studies. *Int J Cancer* 109: 125-131, 2004

Bromen K, Pohlabein H, Jahn I, Ahrens W, Jockel KH: Aggregation of lung cancer families: Results from a population-based case-control study in Germany. *Am J Epidemiol* 152: 497-505, 2000

Brown K: Respiratory Health Effects of Passive Smoking: Lung Cancer and Other Disorders (eds Beyard S, Jinot J, Koppikar AM) Chpt 6, 1-29 (Environmental Protection Agency, Washington DC, USA 1992)

Brownson RC, Alavanja MC, Caporaso N, et al: Family history of cancer and risk of lung cancer in lifetime non-smokers and long-term ex-smokers. *Int J Epidemiol* 26: 256-263, 1997

Brownson RC, Alavanja MC, Caporaso N, Simoes EJ, Chang JC. Epidemiology and prevention of lung cancer in nonsmokers. *Epidemiol Rev* 20: 218-236, 1998

Cancer Facts and Figures 2009, American Cancer Society, Inc. 2009

Chou P, Kuo HS, Chen CH, Lin HC: Characteristics of non-participants and reasons for non-participation in a population survey in Kin-Hu Kinmen. *Eur J Epidemiol* 13:195-200, 1997

Coleman MP, Esteve J, Demieka P, et al: Trends in cancer incidence and mortality. Lyon, France: International Agency for Research on Cancer, 1993

Copper MP, Braakhuis BJ, de Vries N, van Dongen GA, Nauta JJ, Snow GB: A panel of biomarkers of carcinogenesis of the upper aerodigestive tract as potential intermediate endpoints in chemoprevention trials. *Cancer (Phila.)* 71: 825-830, 1993

- Cote ML, Kardia SL, Wenzlaff AS, Ruckdeschel JC, Schwartz AG: Risk of lung cancer among white and black relatives of individuals with early-onset lung cancer. *JAMA* 293(24): 3036-3042, 2005
- Cronin K, Gail MH, Zou Z, et al: Validation of a model of lung cancer risk prediction among smokers. *J Natl Cancer Inst* 98: 637-640, 2006
- Cross FT: Invited Commentary: residential radon risks from the perspective of experimental animal studies. *Am J Epidemiol* 140: 333-339, 1994
- Deng L, Kimmel M, Foy M, Spitz M, Wei Q, Gorlova O: Estimation of the effects of smoking and DNA repair capacity on the coefficients of a carcinogenesis model for lung cancer. *Int J Cancer* 124: 2152-8, 2009
- Deng L: Modelling Carcinogenesis in Lung Cancer: Taking Genetic Factors and Smoking Factor into Account. Doctoral Dissertation. Houston: Rice University, 2005
- Dibble R, Langeburg W, Blair S, Ward J, Akerly W: Natural history of non-small cell lung cancer in non-smokers. *J Clin Oncol* 23: 7252, 2005
- Doll R: Atmospheric pollution and lung cancer. *Environ Health Perspect* 22:23-31, 1978
- Doll R, Hill AB: Smoking and carcinoma of the lung. *BMJ* 2: 739-748, 1950
- Doll R, Peto R: The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst* 334: 1150-1155, 1981
- Early lung cancer detection: summary and conclusions. *Am Rev Respir Dis* 130: 565-570, 1984
- Etter JF, Perneger TV: Analysis of non-response bias in a mailed health survey. *J Clin Epidemiol* 50:1123-8, 1997
- Etzel CJ, Amos CI, Spitz MR: Risk for smoking-related cancer among relatives of lung cancer patients. *Cancer Research* 63: 8531-8535, 2003
- Flehinger BJ, Kimmel M: The natural history of lung cancer in a periodically screened population. *Biometrics*: 43:127-44, 1987
- Franklin WA, Gazdar AF, Haney J, Wistuba II, La Rosa FG, Kennedy T, Ritchey DM, Miller YE: Widely dispersed p53 mutations in respiratory epithelium. A novel mechanism for field carcinogenesis. *J Clin Invest.* 100:2133-2137, 1997
- Friberg L, Cederlof R: Late effects of air pollution with special reference to lung cancer. *Environ Health Perspect* 22: 45-66, 1978

Froom P, Melamed S, Kristal-Boneh E, Benbassat J, Ribak J: Healthy volunteer effect in industrial workers. *J Clin Epidemiol* 52:731-5, 1999

Gorlova OY, Zhang Y, Schabath MB, et al: Never smokers and lung cancer risk: A case-control study of epidemiological factors. *Int J Cancer* 118: 1798-1804, 2006

Gottschall EB: Occupational and environmental exposures. *J Thorac Imaging* 17: 189-197, 2002

Hara M, Sasaki S, Sobue T, Yamamoto S, Tsugane S: Comparison of cause-specific mortality between respondents and nonrespondents in a population-based prospective study: ten-year follow-up of JPHC Study Cohort I. *Japan Public Health Center. J Clin Epidemiol* 55:150-6, 2002

Hasegawa M, Sone S, Takashima S, Li F, Yang ZG, Maruyama Y, Watanbe T: Growth rate of small lung cancers detected on mass screening. *British Journal of Radiology* 73: 1252-1259, 2000

Hazelton WD, Luebeck EG, Heidenreich WF, Moolgavkar SH: Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model. *Radiation Research* 156: 78-94, 2001

Hazelton WD, Clements MS, Moolgavkar SH: Multistage carcinogenesis and lung cancer mortality in three cohorts. *Cancer Epidemiology, Biomarkers, and Prevention* 14(5): 1171-1181, 2005

Hazelton WD, Moolgavkar SH, Curtis SB, Zielinski JM, Ashmore JP, Krewski D: Biologically based analysis of lung cancer incidence in a large Canadian occupational cohort with low-dose ionizing radiation exposure, and comparison with Japanese atomic bomb survivors. *Journal of Toxicology and Environmental Health, Part A* 69: 1013-1038, 2006

Heidenreich WF, Luebeck EG, Moolgavkar SH: Some properties of the two-mutation clonal expansion model. *Risk Analysis* 17(3): 391-399, 1997

Heidenreich WF, Jacob P, Paretzke HG: Exact solution of the clonal expansion model and their application to the incidence of solid tumors of atomic bomb survivors. *Radiat Environ Biophys* 36: 45-58, 1997

Heidenreich WF, Luebeck EG, Moolgavkar SH: Some Properties of the Hazard Function of the Two-Mutation Clonal Expansion Model. *Risk Analysis* 17(3): 391-399, 1997

Heidenreich WF, Jacob P, Paretzke HG, Cross FT, Dagle GE: Two-step model for the risk of fatal and incidental tumors in rats exposed to radon. *Radiation Research* 151: 209-217, 1999

Heidenreich WF, Wellmann J, Jacob P, Wichmann HE: Mechanistic modeling in large case-control studies of lung cancer risk from smoking. *Statistics in Medicine* 21: 3055-3070, 2002

Henschke CI, McCaulery DI, Yankelovitz DF, et al: Early lung cancer action program: overall design and findings from baseline screening. *Lancet* 354: 99-105, 1999

Henschke CI, Naidich DP, Yankelovitz DF et al: Early lung cancer action project: initial findings on repeat screening. *Cancer* 92: 153-159, 2001

Hudmon KS, et al: Identifying and recruiting health control subjects from a managed care organization: a methodology for molecular epidemiological case control studies of cancer. *Cancer Epidemiol Biomarkers Prev* 6: 565-571, 1997

International Agency for Research on Cancer (IARC). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans and their Supplements: A complete list: Tobacco Smoking Volume 38. 1986

International Agency for Research on Cancer (IARC). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans and their Supplements: A complete list: Involuntary Smoking Volume 83. 2002

The International Early Lung Cancer Action Program Investigators. Survival of patients with stage I lung cancer detected on CT screening. *N Engl J Med* 355: 1763-1771, 2006

International Early Lung Cancer Action Program Investigators (IELCAP): Henschke CI, Miettinen YR: Women's susceptibility to tobacco carcinogens and survival after diagnosis of lung cancer. *JAMA* 296(2): 180-184, 2006

Kendall DG: Birth and death processes and the theory of carcinogenesis. *Biometrika* 47: 13-21, 1960

Koo LC, Ho JH, Lee N: An analysis of some risk factors for lung cancer in Hong Kong. *Int J Cancer* 35: 149-155, 1985

Kopp-Shneider A, Portier CJ, Sherman CD: The exact formula for tumour incidence in the two-stage model. *Risk Analysis* 14: 1079-1080, 1994

Levin ML, Goldstein H, Gerhardt PR. Cancer and tobacco smoking: a preliminary report. *JAMA* 143: 336-338, 1950

Lindsted KD, Fraser GE, Steinkohl M, Beeson WL: Healthy volunteer effect in a cohort study: temporal resolution in the Adventist Health Study. *J Clin Epidemiol* 49:783-90, 1996

Littman AJ, Thornquist MD, White E, et al: Prior lung disease and risk of lung cancer in a large prospective study. *Cancer Causes Control* 15: 819-827, 2004

Lubin JH et al: A Joint Analysis of 11 Underground Miner Studies. National Institutes of Health, Bethesda, MD, USA, 1994

Luebeck EG, Heidenreich WF, Hazelton WF, Paretzke HG, Moolgavkar SH: Biologically based analysis of the data for the Colorado Uranium Miners Cohort: age, dose and dose-rate effects. *Radiation Research* 152: 339-351, 1999

Marks F, et al: Tumor promotion as a target of cancer prevention. *Recent Results Cancer Res* 174: 37-47, 2007

Matakidou A, Eisen T, Houlston RS: Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer* 93(7): 825-833, 2005

Mayne ST, Buenconsejo J, Jenerich DT: Familial cancer history and lung cancer risk in United States nonsmoking men and women. *Cancer Epidemiology, Biomarkers, and Prevention* 8: 1065-1069, 1999

Mayne ST, Buenconsejo J, Jenerich DT: Previous lung disease and risk of lung cancer among men and women nonsmokers. *Am J Epidemiol* 149: 13-20, 1999

Meza R, Hazelton WD, Colditz GA, Moolgavkar SH: Analysis of lung cancer incidence in the nurses' health and health professionals' follow-up studies using a multistage carcinogenesis model. *Cancer Causes Control: Epub* December 2007

Moolgavkar SH, Venzon DJ: Two-event models for carcinogenesis: Incidence curves for childhood and adult tumours. *Mathematical Biosciences* 47: 55-77, 1979

Moolgavkar SH, Luebeck G: Two-event model for carcinogenesis: Biological, mathematical, and statistical considerations. *Risk Analysis* 10(2): 323-341, 1990

Moore KA, Mery CM, Jaklitsch MT, Estocin AP, Bueno R, Swanson SJ, Sugarbaker DJ, Lukanich JM: Menopausal effects on presentation, treatment, and survival of women with non-small cell lung cancer. *Ann Thorac Surg* 76: 1789-1795, 2003

Muscat JE, Wynder L: Lung cancer pathology in smokers, ex-smokers and never smokers. *Cancer* 88: 1-5, 1995

National Cancer Institute (NCI): Surveillance Epidemiology and End Results (SEER). Division of Cancer Control and Population Sciences. www.seer.cancer.gov

National Institutes of Health (NIH), National Cancer Institute: Smoking and Tobacco Monograph 10: Health Effects of Exposure to Environmental Tobacco Smoke. 1999

National Institutes of Health (NIH): What you need to know about lung cancer.
Publication No. 07-1553

National Research Council (NRC), Committee on Passive Smoking: Environmental Tobacco Smoke: Measuring Exposures and Assessing Health Effects. 1986

Neuberger JS, Field RW: Occupation and lung cancer in nonsmokers. *Rev Environ Health* 18: 251-267, 2003

Neyman J, Scott EL: Statistical aspects of the problem of carcinogenesis. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics, and Probability* 4: 707-719, 1967

Nordquist LT, Simon GR, Cantor A, Alberts WM, Bepler G: Improved survival in never smokers vs current smokers with primary adenocarcinoma of the lung. *Chest* 126: 347-351, 2004

Patel JD, Bach PB, Kris MG: Lung cancer in US women: a contemporary epidemic. *JAMA* 291: 1763-1768, 2004

Pinsky PF, Miller A, Kramer BS, Church T, Reding D, Prorok P, Gelmann E, Schoen RE, Buys S, Hayes RB, Berg CD: Evidence of a healthy volunteer effect in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. *Am J Epidemiol* 164: 874-881, 2007

Prescott E, Olser M, Anderson PK, Hein HO, Borch-Johnsen K, Lange P, Schnohr P, Vestbo J: Mortality in women and men in relation to smoking. *Int J Epidemiol* 27: 27-32, 1998

Proctor RN: *The nazi war on cancer*. Princeton, NJ: Princeton University Press, 1999

Qiao Y, Spitz MR, Guo Z, Hadeyati M, Grossman L, Kramer KH, Wei Q: Rapid assessment of repair of ultraviolet DNA damage with a modified host-cell reactivation assay using a luciferase reporter gene and correlation with polymorphisms of DNA repair genes in normal human lymphocytes. *Mutat Res* 509: 165-174, 2002

Ries LAG, Miller BA, Hankey BF, et al: *Cancer statistics review, 1973-1988*. Bethesda, MD: US Government Printing Office, 1991

Risch HA, Howe GR, Jain M, Burch JD, Holowaty EJ, Miller AB: Are female smokers at higher risk for lung cancer than male smokers? A case-control analysis by histological type. *Am J Epidemiol* 138: 281-293, 1993

Schwartz AG, Siegfried JM, Weiss L: Familial aggregation of breast cancer with early onset lung cancer. *Genet Epidemiol* 17(4): 274-284, 1999

- Schwartz AG, Yang P, Swanson GM: Familial risk of lung cancer among nonsmokers and their relatives. *Am J Epidemiol* 144(6): 554-562, 1996
- Selikoff IJ, Churg J, Hammond EC: Asbestos exposure and neoplasia. *JAMA* 188: 22-26, 1964
- Sellers TA, Ooi WL, Elston RC, Chen VW, Bailey-Wilson JE, Rothschild H: Increased familial risk for non-lung cancer among relatives of lung cancer patients. *Am J Epidemiol* 126(2): 237-246, 1987
- Shen H, Spitz MR, Qiao Y, Guo Z, Wang LE, Bosken CH, Amos CI, Wei Q: Smoking, DNA repair capacity and risk of non-small cell lung cancer. *Int J Cancer* 107: 84-88, 2003
- Shimizu H, Tominaga S, Nishimura M, Urata A: Comparison of clinico-epidemiological features of lung cancer patients with and without a history of smoking. *Jpn J Clin Oncol* 14: 595-600, 1984
- Sikkink SK, et al: In-depth analysis of molecular alterations within normal and tumour tissue from an entire bronchial tree. *Int J Oncol* 22: 589-559, 2003
- Slaughter DP: The multiplicity of origin of malignant tumors. *Internation Abstracts of Surgery* 79(2): 89-98, 1944
- Slaughter DP, Southwick HW, Smejkal W: "Field Cancerization" in oral stratified squamous epithelium. *Cancer (Phila.)* 6: 963-968, 1959
- Sobue T, Yamamoto S, Hara M, Sasazuki S, Sasaki S, Tsugane S: JPHC Study Group. Japanese Public Health Center: Cigarette smoking and subsequent risk of lung cancer by histological type in middle-aged Japanese men and women: the JPHC study. *Int J Cancer* 99: 245-251, 2002
- Sozzi G, Miozzo M, Pastorini U, Pilotti S, Donghi R, Giarola M, De Gregorio L, Manenti G, Radice P, Minoletti F, Della Porta G, Pierotti MA: Genetic evidence for an independent origin of multiple preneoplastic and neoplastic lung lesions. *Cancer Research* 55: 135-140, 1995
- Subramanian J, Govindan R: Lung cancer in never smokers: a review. *Journal of Clinical Oncology* 25(5): 561-570, 2007
- Subramanian J, Velcheti V, Gao F, Govindan R: Presentation and stage-specific outcomes of lifelong never-smokers with non-small cell lung cancer (NSCLC). *Journal of Thoracic Oncology* 2(9): 827-830, 2007
- Sun S, Schiller JH, Gazdar AF: Lung cancer in never smokers- a different disease. *Nature* 7: 778-789, 2007

Tan WY: Stochastic models of carcinogenesis. Marcel Decker. New York 1991

Texas Cancer Registry: Texas Cancer Information. Cancer Epidemiology and Surveillance Branch. www.texascancer.info

Thomson CA, Harris RB, Craft NE, Hakim IA: A cross-sectional analysis demonstrated the healthy volunteer effect in smokers. *Journal of Clinical Epidemiology* 58: 378-382, 2005

Thun MJ, Myers DG, Day-Lally C, Myers D, Calle EE, et al: Trends in tobacco smoking and mortality from cigarette use in Cancer Prevention Studies I (1959 through 1965) and II (1982 through 1988). In: National Cancer Institute, Smoking and Tobacco control, monograph 8: Changes in cigarette-related disease risks and their implication for prevention and control, 1997

Tockman MS: Other host factors and lung cancer susceptibility. In: Samet JM, ed. *Epidemiology of lung cancer*. New York, NY: Marcel Dekker: 397-412, 1994

Toh CK et al: Never-smokers with lung cancer: epidemiological evidence of a distinct disease entity. *J Clin Oncol* 24: 2245-2251, 2006

Unger M: A pause, progress and reassessment in lung cancer screening: *N Engl J Med* 355: 17, 2006

U.S. Department of Health and Human Services: The Health Consequences of Involuntary Smoking: A Report of the Surgeon General, 1979

U.S. Department of Health and Human Services: Women and Smoking: A Report of the Surgeon General, 2001

U.S. Department of Health Education, and Welfare (DHEW): Smoking and health: a report of the Advisory Committee to the Surgeon General. DHEW-Public Health Service Publication No. 1103. Washington, DC: US Government Printing Office, 1964

U.S. Environmental Protection Agency. Respiratory Health Effects of Passive Smoking. 1992

Usuda K, Saito Y, Sagawa M, Soto M, Kanma K, Takahashi S, Endo C, Chen Y, Sakurda A, Fujimura S: Tumor doubling time and prognostic assessment of patients with primary lung cancer. *Cancer* 74: 2239-2244, 1994

Vineis P, Airoidi L, Vegelia P, et al: Environmental tobacco smoke and risk of respiratory cancer and chronic obstructive pulmonary disease in former smokers and never smokers in the EPIC prospective study. *BMJ* 330: 277, 2005

Wakelee HA, Chang ET, Gomez SL, Keegan TH, Feskanich D, Clarke CA, et al: Lung cancer incidence in never smokers. *J Clin Oncol* 25(5): 472-478, 2007

Wei Q, Cheng L, Amos CI, Wang LE, Guo Z, Hong WK, Spitz MR: Repair of tobacco carcinogen-induced DNA adducts and lung cancer risk: a molecular epidemiologic study. *J Natl Cancer Inst* 92(21): 1764-1772, 2000

Wingo PA, Ries LA, Giovino GA, et al: Annual report to the nation on the status of cancer, 1973-1996, with a special section on lung cancer and tobacco smoking. *J Natl Cancer Inst* 91: 675-690, 1999

Wu AH, Fontham ET, Reynolds P, et al: Previous lung disease and risk of lung cancer among lifetime nonsmoking women in the United States. *Am J Epidemiol* 141: 1023-1032, 1995

Wu AH, Fontham ET, Reynolds P, et al: Family history of cancer and risk of lung cancer among lifetime nonsmoking women in the United States. *Am J Epidemiol* 143: 535-542, 1996

Wu X, et al: p53 Genotypes and Haplotypes Associated with Lung Cancer Susceptibility and Ethnicity. *J Natl Cancer Inst* 94: 681-690, 2002

Wynder EL, Graham EA: Tobacco smoking as a possible etiological factor in bronchiogenic carcinoma: a study of six hundred and eighty-four proven cases. *JAMA* 143: 329-336, 1950

Zang EA, Wynder EL: Difference in lung cancer risk between men and women: examination of the evidence. *J Natl Cancer Inst* 88: 183-192, 1996

Zheng Q: On the exact hazard and survival functions of the MVK model. *Risk Analysis* 14: 1081-1084, 1994