RICE UNIVERSITY

Modeling Carcinogenesis in Lung Cancer: Taking Genetic Factors and Smoking Factor into Account

by

Li Deng

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

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Marek Kimmel, Chair and Co-director
Professor of Statistics

James R. Thompson
Noah Harding, Professor of Statistics

Steven J. Cox
Professor of Computational and Applied Mathematics

Olga Y. Gorlova, Co-director
Assistant Professor of Epidemiology, U.T.
M.D. Anderson Cancer Center

HOUSTON, TEXAS

AUGUST, 2005
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

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

Modeling Carcinogenesis in Lung Cancer: Taking Genetic Factors and Smoking Factor into Account

by

Li Deng

The goal of my thesis is to assess the impacts of cigarette smoking and genetic susceptibility on the onset lung cancer and to compute the age-specific probability of developing lung cancer given risk factor levels. The improvement in predicting the chance of having lung cancer at certain age will enhance physicians' capability to design a sensible screening strategy for early tumor detection in a high-risk population. This is the only way to reduce the mortality rate since no effective treatment or cure is available for advanced lung cancer at this time. The evaluation of the effects of these two risk factors proceeds through parameter estimation in the framework of the two-stage clonal expansion (TSCE) model applied to case-control study data. The TSCE model describes carcinogenesis as transitions from normal cells to slightly abnormal cells and to cancerous cells. Our data analysis indicates that smoking enhances the
proliferation rate while both smoking and genetic susceptibility affect initiation and malignancy transformation rates. The data suggests that there might be a mechanism difference in the development of lung cancer for non-smokers and for smokers. Besides predicting survival rates, I rigorously prove the non-identifiability theorem for the TSCE model in the piecewise constant case and derive a new algorithm of calculating the survival function for a 3-stage and 2-path stochastic model. This 3-stage and 2-path model has two new features: it consists of two stages instead of one for abnormal cells, where one stage is more advanced than the other, and it includes two paths connecting normal cells to cancerous cells. The test of the new model on Texas cancer data shows a very good fit. Such efforts in developing models that incorporate new findings will lead to a better understanding of the mechanism of carcinogenesis and eventually to the development of drugs to treat cancer.
Acknowledgements

First of all, I would like to thank my thesis directors Drs. Marek Kimmel and Olga Y. Gorlova for their guidance, numerous corrections and conversations that significantly contributed to the final form of this thesis. This author would also like to thank the other committee members, Dr. James R. Thompson and Dr. Steven J. Cox, for their helpful suggestions and service. I am indebted to Dr. Keith A. Baggerly for his numerous tips on research and statistical software. I also thank American Cancer Society and M.D. Anderson Cancer Center Epidemiology Department for releasing their data to us. I am also grateful to Dr. Katherine Ensor, Ms. Diane Brown, Ms. Leticia Gonzales, Dr. Gary Rosner, Dr. Dennis Cox, Dr. Javier Rojo and many other faculty and staff members of the Department of Statistics for their help during my graduate study in this department. I would like to acknowledge my roommate Dr. Gretchen A. Fix for helping me to improve my English and my officemate Pawel W. Paszek for his constant encouragement when I was stuck in the research.

This work is mainly supported by the Keck pre-doctoral fellowship and CISNET lung cancer grant.
Contents

Abstract ii

Acknowledgements iv

List of Figures ix

List of Tables xiv

1 Introduction 1

1.1 Biological background of cancer ......................... 4

1.1.1 Definition of cancer ............................... 4

1.1.2 Development of cancer ............................... 5

1.1.3 Dying of cancer .................................. 6

1.2 Lung cancer facts .................................... 7

1.3 Environmental exposure and genetic susceptibility in lung cancer carcinogenesis ........................................ 10

1.4 Overview of the thesis .................................. 11
2 Review of carcinogenesis models

2.1 Nording and Armitage (N-A) multi-stage model .......................... 14
2.2 Pompei and Wilson (P-W) beta model ................................. 17
2.3 Frank’s model ..................................................................... 18
2.4 Two-stage clonal expansion model ........................................... 21
  2.4.1 Constant parameters in the TSCE model ......................... 23
  2.4.2 Piecewise constant parameters in the TSCE model .............. 25
  2.4.3 The number of identifiable parameters in the piecewise case is
        $4K - 1$ ............................................................................... 26
2.5 Multistage clonal expansion models ......................................... 37

3 Alternative carcinogenesis models .............................................. 38

3.1 Three-stage two-path stochastic model ..................................... 39
3.2 Cumulative density function (CDF) of $T_c$ ................................. 39
3.3 Numerical computation of the CDF of $T$ .................................. 43
3.4 Tan’s multi-pathway model and approximation formula ............. 46
3.5 Fitting the three-stage two-path model to the Texas lung cancer data 50

4 Evaluating the effects of smoking and genetic susceptibility ........ 54

4.1 Summary of prospective data and case-control data .................. 55
  4.1.1 Cancer Prevention Study II data ....................................... 56
  4.1.2 M.D. Anderson case-control study ................................... 58
4.1.3 Exclusion of former smokers’ data ........................................ 59
4.2 Overall estimation strategy .................................................. 62
4.3 Random noise and the trade-off between \( \nu \) and \( \mu \) ............. 63
  4.3.1 Two optimization functions and the logit transformation for
      parameters ........................................................................... 64
  4.3.2 Least Squares Estimation (LSE) and the objective function .... 65
  4.3.3 Response function ............................................................ 65
  4.3.4 Estimates of \( \mu \) and \( \nu \) in the absence of random noise .... 66
  4.3.5 Trade-off between parameters \( \mu \) and \( \nu \) in the presence of random
      noise .................................................................................. 67
4.4 Estimation of baseline biological parameters based on never smokers
      data ....................................................................................... 69
  4.4.1 Estimation based on a mixed dataset .................................... 70
  4.4.2 Estimation of the effect of DRC on never smokers lung cancer
      development ........................................................................... 78
4.5 Current smoker parameter estimation ........................................ 79

5 Conclusions and discussions ...................................................... 98
  5.1 Conclusions ............................................................................ 98
  5.2 Problems and future research direction ..................................... 103

Bibliography .............................................................................. 112
Appendix

A Assumptions and PGF derivation of hazard function

A.1 Recursive formula for computing survival and hazard function for piece-wise constant case
List of Figures

1.1 Age-adjusted Cancer Death Rates: Male by sites, U.S. 1930-2000.
   Copy from cancer statistics 2003 ........................................ 7

1.2 Age-adjusted Cancer Death Rates: Female by sites, U.S. 1930-2000.
   Copy from cancer statistics 2003 ........................................ 8

2.1 Stage transitions in Armitage and Nording multi-stage model. NC:
   normal stem cells; IC: intermediate/initiated cells; MC: malignant/cancerous
   cells ................................................................. 15

2.2 Uniform drop of cancer incidence among high age people, from paper
   by Pompei (2001a) ....................................................... 19

2.3 Uniform drop of tumor incidence among old mice, from paper by Pom-
   pei (2001b) ........................................................... 20

2.4 Two-stage clonal expansion model paradigm. NC: normal stem cells;
   IC: intermediate/initiated cells; MC: malignant cells/cancerous cells 22

3.1 Three-stage transition paradigm ......................................... 39
3.2 Comparison of the hazard rate with different step sizes using Deng’s numerical algorithm (Scenario I). In the plot, alpha is the same as $p$ in the theorem. The parameter $p_0$ is the probability for a single cell to jump in one year, $\text{growb}$ is the net growth rate (same as $\alpha - \beta$ in the theorem) for the intermediate cells and $\nu$ is the first transition rate (from normal cell to intermediate cell). ........................................... 46

3.3 Comparison of the hazard rate with different step sizes using Deng’s numerical algorithm (Scenario II). In the plot, alpha is the same as $p$ in the theorem. The parameter $p_0$ is the probability for a single cell to jump in one year, $\text{growb}$ is the net growth rate (same as $\alpha - \beta$ in the theorem) for the intermediate cells and $\nu$ is the first transition rate (from normal cell to intermediate cell). ........................................... 47

3.4 Tan’s multistage and multipath model paradigm .............................. 49

3.5 Comparison of the hazard rates using Tan’s approximation and Deng’s algorithm ................................................................. 50

3.6 Comparison of the hazard rates using Tan’s approximation and Deng’s algorithm ................................................................. 51

3.7 Comparison of the hazard rates using Tan’s approximation and Deng’s algorithm. ................................................................. 51

3.8 Comparison of the hazard rates for TSCE using Tan’s approximation and the closed-form formula. ........................................ 52
3.9 Nonlinear Regression Estimation (TSCE model) for female whites 1996-2000 .................................................. 52

3.10 Nonlinear Regression Estimation (3-stage 2-path) for female whites 1996-2000 .................................................. 53

4.1 SEER, CPS-II and Texas Lung Cancer data comparison. Never and current smokers mortality comes from CPS-II. The former smokers’ incidence rate is computed by subtracting never and current smokers mortality rates from the combined Texas incidence rates. .................. 61

4.2 Estimation properties of $\mu$ and $\nu$ ........................................... 69

4.3 Estimates of parameters when missing incidence data and noise are both present, $a_1 = a_2$ assumption is enforced. The estimates of $a_1$ and $a_2$ are nearly zero. The two lower plots are based on 10 runs. .............. 74

4.4 Estimates of parameters with more complete data and noise, $a_1 = a_2$ assumption is enforced. The estimates of $a_1$ and $a_2$ are nearly zero. The two lower plots are based on 10 runs. ...................... 76

4.5 Density of estimates of $a_1$ and $a_2$, true $a_1 = 1, a_2 = 0.3$ .............. 77

4.6 Dependence between estimates of $a_1$ and $a_2$ ............................... 77

4.7 Fit of the TSCE model for current female smokers 20 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations. ...................... 86
4.8 Fit of the TSCE model for current female smokers 40 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations. .......................... 87

4.9 Fit of the TSCE model for current female smokers 20 cig/day and 40 cig/day using response function 1. .......................... 88

4.10 Fit of the TSCE model for current male smokers 20 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations. .......................... 89

4.11 Fit of the TSCE model for current male smokers 40 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations. .......................... 90

4.12 Fit of the TSCE model for current male smokers 20 cig/day and 40 cig/day using response function 1. .......................... 91

4.13 Fit of the TSCE model for current female smokers 20 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations. .......................... 92

4.14 Fit of the TSCE model for current female smokers 40 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations. .......................... 93

4.15 Fit of the TSCE model for current female smokers 20 cig/day and 40 cig/day using response function 2. .......................... 94
4.16 Fit of the TSCE model for current male smokers 20 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations. .......................... 95

4.17 Fit of the TSCE model for current male smokers 40 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations. .......................... 96

4.18 Fit of the TSCE model for current male smokers 20 cig/day and 40 cig/day using response function 2. .......................... 97

5.1 Evolutionary path for the first rise of a cancerous cell (different paths for different individuals) .......................... 107

5.2 Simulation based incidence rates built on the beta-shifted model. Red dots are observed Texas lung cancer incidence rates and blue dots are simulated incidence rates based on 10000 samples. .......................... 109

5.3 Hazard function for the normal distribution with normal censoring .......................... 110

5.4 Hazard function for the double exponential distribution .......................... 111
List of Tables

4.1 LSE with two optimization functions

4.2 LSE for mixed incidence data without noises

4.3 LSE with DRC infor.

4.4 Estimates for never smokers

4.5 Distribution of initiation age

4.6 Distribution of smoking intensity

4.7 Ratio between the number of people with poor DRC and good DRC

4.8 LSE for smokers cleaned data response function 1

4.9 LSE for smokers cleaned data response function 2

5.1 Cumulative risk comparison
Chapter 1

Introduction

Human beings have never stopped the battle with diseases since the beginning of human society. Huge improvements in public health and achievements in medical sciences in the first half of the twentieth century have resulted in a dramatically reduced premature death rate, higher quality of life and longer life span. Meanwhile, new medical problems have arisen and await solutions. Among them, AIDS, heart disease and cancer have become the largest health problems for modern society. Cancer as an age-related disease has existed for a long time. However, it did not catch much attention of the public until several decades ago because most cancers occur late in life and relatively few people were able to survive into the “cancer age.” The situation has changed in the past 50 years. According to the data released by the American Cancer Society [1], the mortality of cancer constituted around 23 percent of the total mortality in the U.S. in year 2000, only 6 percent less than heart dis-
eases. Among the three principal death causes, namely, heart diseases, cancer and cerebrovascular diseases, cancer mortality is the only one which has increased since 1950, while the other two have been cut by more than a half.

With the population aging, reducing the gap between research in cancer and the needs of public health is increasingly urgent. Policy makers have already noticed such problem and have allocated a considerable amount of money, effort and resources to cancer research. New promises of biotechnology and computer science make this research field more active than ever. In recent years, researchers not only in biology, but also in other disciplines such as mathematical biosciences, statistics and computer sciences have started to cooperate to try to resolve the problem of cancer.

The mission of cancer research is twofold: prevention and treatment. Good prevention strategies include identifying the causes of cancer, adjusting life style or working conditions to limit the risk factors to the lowest level, adopting sensible screening strategies to high risk populations, and detecting cancer at an early stage. For treatment, we must first understand the key points in carcinogenesis and then design drugs or therapies to achieve the goals of killing cancerous cells and controlling the invasion of tumors. At present, with no effective cure available, prevention is obviously the better choice.

My thesis is focused on evaluating the influence of two risk factors, cigarette smoking and genetic susceptibility, in the development of lung cancer using the two-stage clonal expansion (TSCE) model. The result allows us to see the impact each covari-
ate has on development of lung cancer. Parameters estimated from this model will also enable us to predict an individual’s chance of developing lung cancer at certain age given his/her smoking history, genetic susceptibility and gender. Physicians can utilize such information to design a customized screening strategy for each individual and to catch lung cancer at an early stage. Thus we can reduce the mortality rate of lung cancer and improve the quality of life for cancer patients.

All inferences and predictions on lung cancer incidence among the white population are based on results of data analysis. Our data analysis consists of applying the TSCE model to case-control study and age-specific lung cancer mortality data, estimating coefficients associated with cigarette smoking and genetic factors, and making inference on the influence of each factor on the probability of lung cancer onset. In addition to the application of TSCE to real data, I propose a new numerical algorithm to compute the hazard function of a 3-stage 2-path carcinogenesis model and compare it with an alternative approximation approach. The rigorous proof of the non-identifiability theorem for piecewise constant case is completed in this thesis. Our simplified two or three stage model only can answer few questions at present due to our limited knowledge on cancer. The current models are far from complete. We have also developed a beta-shift model in attempt to create a more general simulation paradigm for carcinogenesis models. There are still numerous fundamental questions for scientists to answer, for instance, what is the number of essential stages in cancer development; which stage is easy to interfere with; what factors are linked to which
stage; and how large is the population variability of the transition rate from one stage to another. The prediction result from this new model is not completely satisfying. But we believe that the effort of building new models with greater flexibility for incorporating new discoveries from biological experiments and clinical research is worthwhile. In the long term, good models can help biologists to hypothesize, test and prove or disprove particular mechanisms of tumor genesis and therefore help find the true mechanism of carcinogenesis and discover the most effective treatments for cancer.

To facilitate further discussion, in the following section I will give a brief introduction to concepts related to cancer.

1.1 Biological background of cancer

1.1.1 Definition of cancer

In general, cancer means malignant tumors. A tumor is the growth of abnormal cells beyond control. A benign tumor proliferates, but it is confined to the original location. In contrast, a malignant tumor has developed a specific capacity to invade neighboring normal tissues and migrate to distant organs through the circulatory or lymphatic systems (metastasis). The tumor at the original location is called the primary tumor in contrast to the secondary tumor appearing at a distant site through metastasis [4]. A benign tumor can be promoted into a malignant tumor
once it gains these features. Conversion in the opposite direction is rarely observed.

1.1.2 Development of cancer

The mechanism of development of cancer is still unclear. Hanahan et al. [6] integrate the most recent findings and define the traits of cancer through six features: "The vast catalog of cancer cell genotype is a manifestation of six essential alternations in cell physiology that collectively dictate malignant growth: self-sufficiency in growth signals, insensitivity to growth-inhibitory (anti-growth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis." Their views of the characteristics of cancer represent the main stream among cancer researchers.

From the viewpoint of evolution, carcinogenesis can be described as abnormal cells' struggle for survival and reproduction. The first step in the process, tumor initiation, occurs when normal stem cells become slightly abnormal and start to divide slowly. The substance that triggers this event is called an initiator. Agents including radiation, chemicals, and viruses can serve as initiators. In the presence of a promoter, a substance which stimulates the growth of initiated cells, slightly abnormal cells can reproduce and form a population of initiated cells for further selections. These cells could regress or even disappear for various reasons. But sometimes, tumor progression continues as additional genetic alterations occur within the initiated cells population. Some cells gain selective advantages such as grow-
ing faster, escaping from apoptosis, building up their own blood vessels for nutrition (angiogenesis) and invading other organs (metastasis). By passing these mutations to their progeny, these winner cells out-reproduce other cells and become dominant within the tumor population. Throughout the development of cancer, initiators and promoters are indispensable elements. Promoters must be applied to subjects, previously exposed over a period time to initiators, to guarantee the occurrence of cancer. An agent which can act both as an initiator and a promoter is called a complete carcinogen. For instance, high dosage of radiation is a complete carcinogen for various cancers. As tumors develop into the late stage, for instance, the metastasis stage, they have become aggressive enough to kill their hosts quickly.

1.1.3 Dying of cancer

Advanced cancer including metastasized tumors can kill the host in various ways, while benign tumor or early stage cancer is much less likely to kill patient. Cancer kills its victims by replacing functional cells with cancerous cells, disturbing the chemical balance of the body, weakening the immune system or bringing massive damage to essential organs such as the lungs, the liver, bones, or the digestive system. Eventually, patients die from organ failure, infection or complications [32].

Our attention is centered around the development of lung cancer and its causes. In the following section we will review basic facts and history of lung cancer.
1.2 Lung cancer facts

More than 60 years ago, colon, stomach, prostate and breast cancer were the most common cancers. Things have changed since 1950s when the lung cancer incidence rate increased steadily and its mortality rate surpassed those of the other cancers. It has remained the leading cause of cancer death for the past 50 years (see Figures 1.2 and 1.2).

![Cancer Death Rates* for Men, US, 1930-2001](image)

*Age-adjusted to the 2000 U.S. standard population.

Figure 1.1: Age-adjusted Cancer Death Rates: Male by sites, U.S. 1930-2000. Copy from cancer statistics 2003

Today, lung cancer contributes about 13% of cancer diagnoses and 28% of all
Figure 1.2: Age-adjusted Cancer Death Rates: Female by sites, U.S. 1930-2000. Copy from cancer statistics 2003

cancer deaths. In U.S. it was estimated that 160,440 people died from this disease in 2004. Every year more people die from lung cancer than from breast cancer or prostate cancer even though the latter two are the most prevalent cancer types among women and men, respectively. The high mortality rate is not only due to the high incidence rate, but also due to the shortage of effective treatments for advanced lung cancer. In its early stage, lung cancer can be symptomless, so only 15% of patients are diagnosed at this stage. The standard treatment is to surgically remove the tumor
and apply chemotherapy or radiation therapy afterwards. The one-year survival rate is about 50%. Chances for survival past one year is low [1], especially for metastatic cases. Removal of the primary tumor in the lung cannot stop cancer cells at other sites from reproducing and killing the host. Recurrence of the cancer is almost inevitable and patients can die from lung failure, infection, or other complications.

When lung cancer "suddenly" emerged as number one killer, people began to search for the causes. Numerous studies have been carried out worldwide. As early as the 1950s eight case-control studies identified cigarette smoking as the main cause of lung cancer and later cohort studies confirmed this association. The simple fact that among all lung cancer patients, around 85% are current or former smokers, clearly shows such association. Another piece of evidence comes from a study in England and Wales on tobacco consumption from 1931 to 1992. The data demonstrated that the increase of lung cancer mortality rates consistently follows the increase of tar-weighted cigarette consumption per adult for both females and males. The curve of lung cancer incidence is about 20 to 25 years delayed behind the curve of tobacco consumption. Additional chemical analysis indicated that cigarettes contain a collection of chemical compounds, which can act as complete carcinogens and induce cancer in a shorter time. Therefore, scientists conclude that cigarette smoking is largely to blame for lung cancer.

Around 10% to 15% of smokers develop lung cancer during lifetime. However, some nonsmokers get lung cancer in their late 30s or early 40s. These facts im-
ply that some factors other than cigarette smoking also play an important role in lung cancer development. A study carried out at the M.D. Anderson Cancer Center Epidemiology Department indicates that genetic susceptibility is another important factor besides smoking [28]. The study also suggested that a genetic component should be very crucial in the absence of smoking. These results make the study of genetic susceptibility more meaningful since a large proportion of the population is non-smokers.

Our model of studying the interaction of cigarette smoking and genetic factors is based on conjectures proposed by molecular biologists according to their experimental findings.

1.3 Environmental exposure and genetic susceptibility in lung cancer carcinogenesis

The process by which the genetic components and the environmental exposure interact is very complex and it is not completely understood. However, scientists have a rough picture of their roles in inducing mutations. Tobacco contains a variety of carcinogens. These compounds are not carcinogenic until metabolized by cellular enzymes to a reactive derivative. Then these metabolites are detoxified and excreted after a certain period of time. The process of metabolizing the carcinogens includes activation, detoxification, and excretion of the carcinogens, with each step controlled
by a group of genes. Therefore, not only cigarette smoking intensity but also genes involved in the metabolic pathways determine the effective carcinogen exposure intensity and exposure time. Carcinogens may cause DNA adducts or chromosome breaks. The extent of this DNA damage differs among people. The frequency of DNA damage induced by mutagens (carcinogens) is referred to as mutagen sensitivity. DNA damage, if it not repaired in time, may turn on or off particular genes, which are the key control genes in tumor development. In the long time of evolution of life, the defense system has developed a specific mechanism called DNA repair to correct errors and restore DNA information. But this system is not perfect and the efficiency of fixing DNA damage varies from individual to individual. If DNA damage is not corrected in time, the defense system most likely forces the cells with damaged DNA to go to programmed cell death (apoptosis) process. Somehow, some cells escape from both checkpoints and are able to survive and gain abnormal function to reproduce without control. Cells bearing such mutations outgrow the rest. They usually require a next mutation to gain more selective advantage and progress even further.

1.4 Overview of the thesis

In the present chapter (Chapter 1), we provide a brief introduction to the problems we are interested in and to the biological background of cancer, especially lung cancer. In Chapter 2, some popular carcinogenesis models are reviewed with focus on the two-stage clonal expansion model (TSCE) and the complete proof of the non-identifiability
theorem for the piecewise constant setting is given. In Chapter 3, we propose a new algorithm to compute the survival function for a three-stage two-path clonal expansion model based on the TSCE framework. Data analysis combining different data sources and the application of the TSCE model to estimate the parameters associated with the risk factors of interest are presented in Chapter 4. In Chapter 5 we conclude and discuss further research.
Chapter 2

Review of carcinogenesis models

About 140 years ago, a German microscopist, Johannes Mueller, showed that cancers were made up of cells. This discovery began the search for changes that would help to pinpoint the specific differences between normal and cancer cells. Rapid advances in biological technology, particularly in cell and molecular biology, now allow us to try to answer questions which even ten years ago could not be approached. Although there are numerous achievements in cancer research from the biological area, mathematical modeling of cancer development is still in its infancy. The first attempt to build a carcinogenesis model based on mathematical theory was the early 1950s papers by Nording, Armitage and Doll ([22] and [2]). There was rapid growth in carcinogenesis modeling immediately following Armitage's paper. The most important feature of these new works is the clonal expansion of abnormal cells, with each model differing from the others by one or two assumptions. Basing their work on earlier clonal ex-
pansion models, Moolgavkar et al [20] developed a two-stage clonal expansion model. They then applied different cancer data sets to the new model and demonstrated that the model was general enough to produce the exponential increase in the age-specific cancer incidence rate. Most importantly, they used their model in risk assessment to identify risk factors in terms of how these factors accelerate transition and proliferation in cancer development. Most of the current carcinogenesis models can be classified to two categories according to whether or not they explicitly consider the clonal expansion process. In this chapter we will review some typical examples in each group.

2.1 Nording and Armitage (N-A) multi-stage model

The pioneering work in carcinogenesis modeling can be traced back to early 1950s. Nording, Armitage and Doll [2] used a stochastic model to explain the linear relationship between age-specific incidence rates and age on the logarithmic scale. Their model describes tumor evolution as a finite number of stage transitions from a normal cell to a malignant one (figure 2.1).

In the N-A model a constant probability of occurrence of mutation throughout the lifetime is assumed and each mutation is a relatively rare event. This leads to the formula for the age-specific incidence (hazard) rate \( I(t) = at^{k-1} \) or \( \ln I(t) = \ln a + (k-1)\ln t \), where \( k \) is the number of stages and \( a \) indicates effects of various factors representing environmental exposure, genetic susceptibility, and dietary factors. In
Figure 2.1: Stage transitions in Armitage and Nording multi-stage model. NC: normal stem cells; IC: intermediate/initiated cells; MC: malignant/cancerous cells
	his model the slope of the fitted line can be interpreted as the number of transitions. The mathematical basis for this formula is simple. Suppose mutation $i$ occurs in time interval $[0, t]$ with probability $p_i t$, then by the time $t$ the probability that $k - 1$ independent mutations occurred is $\prod_{i=1}^{k-1} (p_i t)$. There are $(k - 1)!$ possible orderings of the mutations and the right-ordered one will have probability $\frac{1}{(k-1)!} (\prod_{i=1}^{k-1} p_i t^{k-1})$. For a more rigorous proof, please see Armitage’s paper [2].

Armitage and Doll [2] examined incidence data in England and Wales in 1950 and 1951, within the age limits of 25 to 74, for various types of cancers. The data were successfully fit by the previous formula, for esophageal, stomach, pancreatic, colon and rectal cancer for both females and males. However, there were clear deviations between the fitted line and observation points for cancer of the lung, bladder, prostate, breast, ovary, cervix uteri and corpus uteri. At that time, it was already realized that cancer incidence was not exclusively due to a mechanism of multiple mutations and that some other factors such as hormonal control of growth might play an independent part. For sex organ cancers, hormonal secretion varies widely during lifetime.
Violations of the basic assumptions of the model probably resulted in poor fits. It was also recognized that cigarette smoking had a great impact on the increase of lung cancer incidence and the authors argued that the deviation might be a consequence of that. Armitage suggested a remedy for the case where the transition probability changes over time. For a detailed argument, see Armitage and Doll [2].

The N-A model enjoys great mathematical simplicity. Thus, numerous scientists later continued to use this model to interpret their experimental data. Some of them slightly modified this theory and started new branches of carcinogenesis models. However, as we pointed out earlier, the multistage model did not incorporate one of the key features in cancer development: clonal expansion. Also, no concrete evidence supported the conclusion that there are 6 or 7 stages in carcinogenesis. However, this was the first theory developed in the mathematical world to model the mechanism of tumor development. It demonstrated the applicability of mathematical models in biological research and established a bridge between the two fields. The N-A model helped scientists to approximate the underlying mechanism of tumorigenesis. The data Armitage and Nording were using truncated out the advanced age group information because at that time little data was available for old people and the authors did not trust those numbers. Almost 50 years later, a group at Harvard modified the N-A model to interpret special phenomena occurring in cancer incidence rates among the advanced age group and proposed new hypotheses of carcinogenesis.
2.2 Pompei and Wilson (P-W) beta model

Many new hypotheses stem from observations based on recently collected datasets. Today, the availability and accumulation of more complete and more accurate data allow us to have a closer look at cancer incidence rates. After observing a “universal” decrease (figure 2.2) in cancer incidence among old people across all cancer types worldwide, Pompei and Wilson [27] proposed another model using the beta distribution to fit cancer incidence curves. Their model essentially is a modification of the N-A model obtained by adding an overturn factor, which results in a new incidence function \( I(t) = (at)^{k-1}(1 - bt) \). The cancer overturn factor \( 1 - bt \) allows the incidence curve to go downward after certain age and thus dramatically improves the goodness of fit. The authors investigated data from lung, colon, stomach, pancreatic, leukemia, breast, prostate and ovarian cancers collected in four different centers around the world. All data fit their model very well. The initial motivation of the Pompei-Wilson model was merely to improve the fit as much as they could. The beta model does not add extra compartments or more complicated pathways to the N-A multistage model. The authors offered some biological explanations which might account for the uniform drop in cancer incidence among old people. The two possibilities they suggested are: first, the mechanism of “cancer extinction” might be due to increasing apoptosis in vivo with the level of accumulated injury related to aging. Second, cell senescence or loss of proliferative ability may be dragging the incidence down. Their interpretation of these mechanisms still needs confirmation from future
biological research. Pompei and Wilson’s model would be more realistic and plausible if after a competing risk adjustment the incidence data still demonstrated a consistent downturn trend among old people. In a later paper [26], they demonstrated that similar incidence curves were observed in induced mice tumor data at all levels of carcinogen exposure (see figure 2.3). This evidence made their hypotheses more convincing, however more investigation will be necessary to verify the mechanism they proposed for human cancer. Determining the turnover age also presents an interesting research topic.

2.3 Frank’s model

Another modification of the N-A model was proposed by a biologist when he was analyzing retinoblastoma and colon cancer incidence. Earlier this year, Steven Frank [5] from Johns Hopkins University gave a simple model to account for the differences between the age-specific incidences of inherited versus sporadic cancers. He utilized the N-A model’s assumptions and derivation in the analysis of colon cancer data and reached the conclusion that the ratio between the age-specific incidence of inherited versus sporadic cancers should be almost linear through the following formula: \( R = \frac{P_{n-1}}{P_{n-2}} = \frac{u!}{n-1} \), where \( P_n \) is the probability that \( n \) transitions happened in a particular order and each event has probability \( u \)t to occur by age of \( t \). In his second example using retinoblastoma cancer, he formulated the ratio between unilateral and bilateral cases as \( \frac{I_u(t)}{I_B(t)} = \nu C(t) \), where \( \nu \) is the mutation rate and \( C(t) \) is the number of cell
Figure 2a-f: Age specific incidence (per 100,000) vs. age for males and females. Beta distribution fits of SEER (Rets S et al. 2000) data for non-gender-specific sites. Parameter values are listed for the Beta function form: \( f(t) = \frac{\alpha t^{\alpha-1} (1-t)^{\beta-1}}{B(\alpha, \beta)} \) where \( B(\alpha, \beta) \) is the Beta function. The fit values are calculated as the fraction of the variance of the observed data which are accounted for by the Beta model with the listed parameter values.

Figure 2.2: Uniform drop of cancer incidence among high age people, from paper by Pompei (2001a)
Figure 2. (a–c) Age-specific mortality (including morbidity) caused by the three most common causes of death by neoplasia for ED91 uncoated control animals and data fit by the beta model. Tests of significant changes show in all cases that the oldest age group (600–1001 days) has significantly lower age-specific mortality than the 600–800 days group, which in turn has significantly higher age-specific mortality than both the 400–600 and the 200–400 days groups. Calculated age-specific incidence for the same tumour sites from data by Sheldon and Greenman (1988) are shown for comparison.

Figure 2.3: Uniform drop of tumor incidence among old mice, from paper by Pompei (2001b)

divisions at age $t$. No specific property of the function $C(t)$ was given. As the author argued in the paper, in the presence of incidence data alone, a simpler model might be
more appealing since there are no experimental data available with which to precisely estimate parameters for intermediate processes. However, since his new formula was proposed for a special dataset, it might be lacking in generality. Moreover, in his colon cancer example, the ratio plot clearly shows a decrease from age 20 to 30 and then a linear increase after 30 in semi-log scale. This does not agree with the linear trend predicted by ut. Frank explained that the initial decrease is due to statistical fluctuation. Furthermore, in the second retinoblastoma case, there are only three data points in the ratio plot. It is nearly impossible to test validity of any model with three sample points. Although the derivation in this paper is not rigorous, it is a good example of biologists' preference for simple models rather than complicated ones. This is one of the reasons why the N-A model still remains popular among biologists 50 years after its publication. Note that Frank's model was designed for ratio type data, so it cannot be used for fitting incidence data alone, or applied to our study.

2.4 Two-stage clonal expansion model

The former three models track transitions in tumorigenesis. However, as seen from experimental data, cancer develops from clonal tumors. The outgrowth of abnormal cells characterizes another key feature of cancer. Based on such motivation, Armitage and Doll [3] developed a two-stage model, with the first stage involving the exponential growth of intermediate cells and the second stage involving the escape from control.
In 1960 Kendall [13] worked out a two-stage theory with the first stage consisting of a subcritical birth and death process (death rate greater than birth rate) and the second stage consisting of a supercritical birth and death process (birth rate greater than death rate) where malignant cells give rise to malignant tumors. Neyman and Scott [21] proposed a similar two-stage model in which the intermediate cells generate clones according to a subcritical birth and death process. Moolgavkar and Venzon [20] developed a two-stage clonal expansion model (see illustration in figure 2.4). This model allows for growth and differentiation of both normal and intermediate cells. Clonal expansion follows a supercritical birth and death process. A set of theories have since been derived from this model. These so-called MVK, or two-stage clonal expansion (TSCE) models, have been successfully used in risk assessments to evaluate the risks of different factors and carcinogens in human cancers. In the rest of this chapter, we will discuss properties of the TSCE model in detail.

Figure 2.4: Two-stage clonal expansion model paradigm. NC: normal stem cells; IC: intermediate/initiated cells; MC: malignant cells/cancerous cells
In the TSCE model, there are three types of cells: normal stem cells, intermediate cells and malignant cells. A single malignant cell will evolve into a malignant tumor with probability one. We denote normal stem cells as NCs, intermediate (initiated) cells as ICs and malignant cells as MCs. For every single NC or IC, the waiting time for it to divide, to die, or to jump to the next stage follows an exponential distribution with particular parameters associated with each process. We assume that the dynamic changes in each cell are independent from those in other cells.

In the general TSCE model framework, the transition rate from the normal cell pool to the intermediate cell pool follows a Poisson process with density \( \nu(t) \), the birth and death rate for ICs are \( \alpha(t) \) and \( \beta(t) \), respectively, and the malignancy transformation rate for a single cell is \( \mu(t) \), where \( t \) denotes the age. In the TSCE model, the survival function is referred as the survival function of the waiting time \( T \) from birth to the rise of the first malignant cell. The time from the rise of the first malignant cell to the death of its host usually is short and can be considered negligible. Therefore in this thesis, we consider the survival function of the waiting time from birth to death (from cancer) to be the same as the survival function of \( T \).

First, let us examine the properties related to the simplest TSCE model.

### 2.4.1 Constant parameters in the TSCE model

When all parameters are constant during a lifetime, the TSCE model is reduced to a homogeneous two-stage model. For some cancers such as colon cancer, stomach
cancer, lung cancer for never smokers and so on, the constant parameters assumption is plausible. In the 1980s, hazard function for this simplest case was computed by numerically solving partial differential equations. Almost fifteen years later, Zheng [34] and Kopp-Schneider et al.[15] independently derived a closed form solution for the survival and hazard functions and published their results in Risk Analysis in back to back. Here we take the result from Zheng’s paper.

**Theorem 2.4.1** In the homogeneous two-stage clonal expansion model, when the number of normal cells is considered constant over time, the hazard function $\lambda(t)$ and survival function $S(t)$ are given by:

$$
\lambda(t) = \frac{\nu(\alpha - \beta - \mu + C)(-\alpha + \beta + \mu + C)(e^{Ct} - 1)}{2\alpha[(-\alpha + \beta + \mu + C) + (\alpha - \beta - \mu + C)e^{Ct}]}
$$

and

$$
S(t) = \left[\frac{2Ce^{0.5(-\alpha + \beta + \mu + C)t}}{(-\alpha + \beta + \mu + C) + (\alpha - \beta - \mu + C)e^{-Ct}}\right]^{\nu/\alpha},
$$

where $C = \sqrt{(\alpha + \beta + \mu)^2 - 4\alpha\beta}$.

In early applications of this TSCE model, researchers noticed that there might be a problem of over-parameterization when using incidence data alone to estimate the four biological parameters, $\alpha$, $\beta$, $\mu$ and $\nu$. In 1996, Heidenreich [10] clearly pointed out that when fitting this model to age-specific incidence data, only three of the four parameters can be determined provided that the incidence curve reaches a plateau. Hanin [7] rigorously proved that three combinations of the four parameters uniquely determine the survival and hazard functions. Heidenreich proposed a traditional
combination of the four parameters and re-derived the formula for constant case as follows:

\[ \lambda(t) = \frac{M(e^{(\gamma+2B)t} - 1)}{\gamma + B(e^{(\gamma+2B)t} + 1)}, \]

where \( M = \mu \nu, \gamma = \alpha - \beta - \mu \) and \( B = (-\gamma + \sqrt{\gamma^2 + 4\alpha \mu})/2 \). These three new parameters have straightforward biological interpretations. The parameter \( \gamma \) is approximately the net proliferation rate, \( M \) is the measure of the first and second transition speeds, and \( B \) is associated with the asymptotic height of the hazard function. As we can see from this formula, it is easy to estimate \( \alpha - \beta \) and \( M \), but hard to tell the absolute values of \( \alpha, \beta, \mu, \) and \( \nu \) even if the two new parameters \( (\gamma, M) \) are known. More properties of the incidence function and parameters can be found in Heidenreich's paper [11]. It can be shown that this formula is equivalent to Zheng's. Later, in our data analysis, we will study age-specific lung cancer incidence and estimate parameters from incidence data and case-control data. Their analytical formula will make our estimation more precise and faster.

### 2.4.2 Piecewise constant parameters in the TSCE model

Although solving for the survival function, \( S(t) \), and the hazard function, \( \lambda(t) \), analytically is impossible when all parameters are time dependent, the solution for the piecewise constant case is tractable. In real life it is practical to treat some age dependent cases as piecewise constant cases. Therefore it is meaningful to derive formulas for the hazard and survival functions under the piecewise constant setting. For in-
stance, the transition rate and proliferation rate can shift after smokers start and quit smoking. In the early 1990s, Tan [29] derived a general formula for this case through the probability generating function (pgf) approach. However, solving partial differential equations (PDEs) is not trivial in this case and numerical approximations must be applied. In 1997, Heidenreich [11] gave a recursive, easily programmable formulas for the first time. These formulas were used for risk assessment of radon, smoking and other environmental factors in lung cancer development. Our data analysis for estimating the effect of cigarette smoking in the development of lung cancer will implement this recipe. Heidenreich’s exact formulas can be found in the Appendix.

In his paper, Heidenreich loosely argued that since the survival function, $S(t)$, can be written as a function of age $t$ and another $4K - 1$ parameters, the number of identifiable parameters is $4K - 1$ when there are $K$ subintervals. His claim is correct. However, his argument is informal. In the next section, we will make the theory of non-identifiability complete by providing a mathematically rigorous proof for Heidenreich’s claim.

2.4.3 The number of identifiable parameters in the piecewise case is $4K - 1$

The formula to compute the survival function is given in Heidenreich’s 1997 [11] paper, in equation (17) on page 393.
\[ \log S(t) = \sum_{i=1}^{k} \frac{\nu_i}{\alpha_i} \log \frac{\widetilde{B}_i - \widetilde{A}_i}{f_i(t_{i-1}, t)} \] (2.1)

The new parameters \( \frac{\nu_i}{\alpha_i}, \frac{\alpha_i - 1}{\alpha_i}, \widetilde{A}_i, \) and \( \widetilde{B}_i \), where \( i = 1, \ldots, K \), completely determine the survival function. However this does not directly establish whether or not the agreement of two survival functions on \( [0, \infty) \) necessarily implies that all \( 4K - 1 \) parameters can be uniquely identified. We will follow Hanin’s proof style and show that the number of identifiable parameters for \( K \) subintervals is \( 4K - 1 \).

Before we start our proof, let us introduce some notation and function definitions from Heidenreich’s paper that will aid in the understanding of our proof. Some of our notation varies from that of Heidenreich. We believe that by rewriting the formulas in a new form, the proof will be easier to follow.

In the following arguments, we consider the piecewise constant case based on the TSCE model with \( K \) subintervals. The positive real line is divided into subintervals \([0, t_1), [t_1, t_2), \ldots, [t_{K-1}, \infty)\) and the four biological parameters \( \alpha(t), \beta(t), \nu(t), \) and \( \mu(t) \) are constant within each subinterval. Let \( \alpha_i, \beta_i, \nu_i, \) and \( \mu_i \) denote their values on the \( i \)th subinterval \([t_{i-1}, t_i]\). New composite parameters are defined as

\[ \widetilde{A}_i = \frac{1}{2}(\gamma_i - \sqrt{\gamma_i^2 + 4\alpha_i \mu_i}), \] (2.2)

\[ \widetilde{B}_i = \frac{1}{2}(\gamma_i + \sqrt{\gamma_i^2 + 4\alpha_i \mu_i}), \] (2.3)

where \( \gamma_i \) is given by

\[ \gamma_i = \alpha_i - \beta_i - \mu_i. \] (2.4)
It is easy to verify that \(A_i < 0\) and \(B_i > 0\). The survival function \(S(t)\) is defined as \(P(T > t)\). It holds that \(S(0) = 1\) and \(S(\infty) = 0\).

For any \(t \in [t_{k-1}, t_k]\), the formula for the logarithm of the survival function, given in equation 2.1, can be written as:

\[
\log S(t) = \sum_{\{i:t_i > t\}} \frac{\nu_i}{\alpha_i} \log \frac{\tilde{B}_i - \tilde{A}_i}{f_i(t_{i-1}, t_i)} + \frac{\nu_k}{\alpha_k} \log \frac{\tilde{B}_k - \tilde{A}_k}{f_k(t_{k-1}, t)}
\]

where

\[
f_i(t_{i-1}, t_i) = (\tilde{y}_i - \tilde{A}_i)e^{\tilde{B}_i(t_{i-1} - t_i)} + (\tilde{B}_i - \tilde{y}_i)e^{\tilde{A}_i(t_{i-1} - t_i)},
\]

and

\[
f_k(t_{k-1}, t) = (\tilde{y}_k - \tilde{A}_k)e^{\tilde{B}_k(t_{k-1} - t)} + (\tilde{B}_k - \tilde{y}_k)e^{\tilde{A}_k(t_{k-1} - t)}, \tilde{y}_k = 0.
\]

The following theorem is the key to solving the identifiability problem. The original idea of this proof is similar to that of Hanin [7], but with suitable modifications. Our result only requires \(x\) to be in a closed interval, rather than on the whole positive real line. This stronger result enables us to decompose the survival function into \(K\) pieces when the biological parameters are constant on closed intervals.

**Theorem 2.4.2** If \(c_1, c_2 > 0, d_1, d_2 > -1\), and

\[
\frac{e^{c_1x} - 1}{e^{c_1x} + d_1} = \frac{e^{c_2x} - 1}{e^{c_2x} + d_2}, x \in [a, b], a > 0
\]

then \(\lambda = 1, c_1 = c_2\) and \(d_1 = d_2\).

**Proof:** Rearrange the equation above and define a new function \(f(x)\)

\[
f(x) = (e^{c_1x} - 1)(e^{c_2x} + d_2) = \lambda(e^{c_2x} - 1)(e^{c_1x} + d_1).
\]
Taking the first, second and third derivatives of the function, we obtain three equations:

\[ f'(x) = (c_1 + c_2)e^{(c_1+c_2)x} + c_1 d_2 e^{c_1 x} - c_2 e^{c_2 x} = \lambda (c_1 + c_2)e^{(c_1+c_2)x} + \lambda c_2 d_1 e^{c_2 x} - \lambda c_1 e^{c_1 x} \]  
(2.5)

\[ f''(x) = (c_1 + c_2)^2 e^{(c_1+c_2)x} + c_1^2 d_2 e^{c_1 x} - c_2^2 e^{c_2 x} = \lambda (c_1 + c_2)^2 e^{(c_1+c_2)x} + \lambda c_2^2 d_1 e^{c_2 x} - \lambda c_1^2 e^{c_1 x} \]  
(2.6)

\[ f'''(x) = (c_1 + c_2)^3 e^{(c_1+c_2)x} + c_1^3 d_2 e^{c_1 x} - c_2^3 e^{c_2 x} = \lambda (c_1 + c_2)^3 e^{(c_1+c_2)x} + \lambda c_2^3 d_1 e^{c_2 x} - \lambda c_1^3 e^{c_1 x}. \]  
(2.7)

Plugging \( x = a \) into equation (2.5) and combining like items, we have

\[ (c_1 + c_2)e^{(c_1+c_2)a}(1 - \lambda) = c_2(\lambda d_1 + 1)e^{c_2 a} - c_1(\lambda + d_2)e^{c_1 a}. \]  
(2.8)

Similarly plugging \( x = a \) into equations (2.6) and (2.7), we get

\[ (c_1 + c_2)^2 e^{(c_1+c_2)a}(1 - \lambda) = c_2^2(\lambda d_1 + 1)e^{c_2 a} - c_1^2(\lambda + d_2)e^{c_1 a} \]  
(2.9)

\[ (c_1 + c_2)^3 e^{(c_1+c_2)a}(1 - \lambda) = c_2^3(\lambda d_1 + 1)e^{c_2 a} - c_1^3(\lambda + d_2)e^{c_1 a}. \]  
(2.10)

We consider two cases for \( \lambda \):
(1) If \( \lambda = 1 \), then equations (2.8) and (2.9) are reduced to

\[
c_2(d_1 + 1)e^{c_2a} - c_1(1 + d_2)e^{c_1a} = 0
\]

(2.11)

\[
c_2^2(d_1 + 1)e^{c_2a} - c_1^2(1 + d_2)e^{c_1a} = 0.
\]

(2.12)

Let \((d_1 + 1)e^{c_2a} = y\) and \((1 + d_2)e^{c_1a} = z\), recalling the condition that \(d_1, d_2 > -1\), \(y\) and \(z\) cannot be zero. The linear equation system has a nonzero solution, so the coefficient matrix has to be singular. That is, its determinant must be zero. This occurs iff \(c_1 = c_2\). Replacing \(c_1\) with \(c_2\) in the equations above, we obtain \(d_1 = d_2\).

(2) If \( \lambda \neq 1 \), then none of the three equations (2.8), (2.9) and (2.10) would equal zero since \(c_1, c_2 > 0\). Obviously, the two ratios, \(\frac{2.9}{2.8}\) and \(\frac{2.10}{2.9}\), are equal, i.e.

\[
\frac{c_2^2(\lambda d_1 + 1)e^{c_2a} - c_1^2(\lambda + d_2)e^{c_1a}}{c_2(\lambda d_1 + 1)e^{c_2a} - c_1(\lambda + d_2)e^{c_1a}} = \frac{c_2^3(\lambda d_1 + 1)e^{c_2a} - c_1^3(\lambda + d_2)e^{c_1a}}{c_2^2(\lambda d_1 + 1)e^{c_2a} - c_1^2(\lambda + d_2)e^{c_1a}} = (c_1 + c_2) \equiv k.
\]

Substituting \((\lambda d_1 + 1)e^{c_2a}\) with \(y\), \((\lambda + d_2)e^{c_1a}\) with \(z\), and assuming the right side of equation (2.10) to be \(\alpha\), the equality of these ratios gives three equations and a system of linear equations with at least one solution:

\[
c_2^3y - c_1^3z = \alpha
\]

(2.13)

\[
c_2^2y - c_1^2z = \frac{\alpha}{k}
\]

(2.14)

\[
c_2y - c_1z = \frac{\alpha}{k^2}
\]

(2.15)

Recall the following results from linear algebra: a linear equation system has a solution(s) iff the rank of the coefficient matrix \(M\) is the same as the rank of the augmented matrix \(N\), where
\[ M = \begin{pmatrix} c_2^3 & -c_1^3 \\ c_2^2 & -c_1^2 \\ c_2 & -c_1 \end{pmatrix}, \]

and

\[ N = \begin{pmatrix} c_2^3 & -c_1^3 & \alpha \\ c_2^2 & -c_1^2 & \frac{\alpha}{k} \\ c_2 & -c_1 & \frac{\alpha}{k^2} \end{pmatrix}. \]

The rank of \( M \) cannot exceed 2. Thus the determinant of \( N \) has to be zero.

We will compute the determinant of \( N \) using matrix determinant properties.

\[
\det(N) = \det \begin{pmatrix} c_2^3 & -c_1^3 & \alpha \\ kc_2^2 & -kc_1^2 & \alpha \frac{1}{k^3} \\ k^2c_2 & -k^2c_1 & \alpha \end{pmatrix}. 
\]

Expanding the determinant at the third row and the third column, we have

\[
\det(N) = \det \begin{pmatrix} c_2^3 - kc_2^2 & -c_1^3 + kc_1^2 & 0 \\ kc_2^2 - k^2c_2 & -kc_1^2 + k^2c_1 & 0 \\ k^2c_2 & -k^2c_1 & \alpha \end{pmatrix} \frac{1}{k^3}. 
\]

After some simplifications, we have \( \det(N) = k^4c_2c_1\alpha(c_1 - c_2)(k - c_1)(k - c_2) \). The condition \( \det(N) = 0 \) implies that \( c_1 = c_2 \) or \( k = c_1 \) or \( k = c_2 \). Since \( c_1 + c_2 = k \)
and $c_1, c_2 > 0$, the last two conditions are impossible. Hence $c_1 = c_2$. Plugging into equations (2.8) and (2.9) and simplifying, we reach the following equalities:

$$4(1 - \lambda)e^{c_1 \alpha} = \lambda(d_1 - 1) + 1 + d_2 \tag{2.16}$$

$$2(1 - \lambda)e^{c_1 \alpha} = \lambda(d_1 - 1) + 1 + d_2 \tag{2.17}$$

These two equations yield $\lambda = 1$, which reduces to case (1). Finally, we conclude that $d_1 = d_2$. ◦

Now we can follow the notation defined earlier in this section and prove the identifiability theorem:

**Theorem 2.4.3** In the TSCE model under the piecewise constant case, suppose that there are $K$ subintervals and two survival functions $S_1(t)$ and $S_2(t)$ specified by formula 2.1. If $S_1(t)$ and $S_2(t)$ are identical on the positive real line, then the equality

$$S_1(t) = S_1(t; \mu_{1i}, \nu_{1i}, \alpha_{1i}, \beta_{1i}) = S_2(t; \mu_{2i}, \nu_{2i}, \alpha_{2i}, \beta_{2i}) = S_2(t), t \geq 0$$

holds if and only if

$$\tilde{A}_{1i} = \tilde{A}_{2i} \tag{2.18}$$

$$\tilde{B}_{1i} = \tilde{B}_{2i} \tag{2.19}$$
\[
\frac{\nu_{1i}}{\alpha_{1i}} = \frac{\nu_{2i}}{\alpha_{2i}} \quad \text{(2.20)}
\]

\[
\frac{\alpha_{1,(i-1)}}{\alpha_{1,i}} = \frac{\alpha_{2,(i-1)}}{\alpha_{2,i}} \quad \text{(2.21)}
\]

where \( i = 1, \ldots, K \) and \( \tilde{A}_i \) and \( \tilde{B}_i \) are defined in equations (2.2) and (2.3).

**Proof:** The proof consists of four parts:

1) The equality of \( S_1(t) = S_2(t) \) on \((0, t_1]\) gives equalities (2.18), (2.19) and (2.20) for \( i = 1 \).

2) The equality of \( S_1(t) = S_2(t) \) on \((t_1, t_2]\) gives equalities (2.18), (2.19) and (2.20) for \( i = 2 \).

3) The continuity of \( S_1(.) \) and \( S_2(.) \) at point \( t_1 \) gives (2.21) for \( i = 2 \).

4) Repeating 1) to 3) on the interval \((t_i, t_{i+1}]\) where \( i = 2, \cdots, K \), yields (2.18), (2.19), (2.20) and (2.21) for \( i = 2, \cdots, K \).

1) Let us focus on the case where the value of \( t \) varies within the first subinterval \([0, t_1)\). In this case equation (2.1) simplifies to:

\[
\frac{\nu_1}{\alpha_1} \log \frac{\tilde{B}_1 - \tilde{A}_1}{f_1(0,t)} = \nu_1 \log(\tilde{B}_1 - \tilde{A}_1) - \frac{\nu_1}{\alpha_1} \log f_1(0,t),
\]

where \( f_1(0,t) = -\tilde{A}_1 exp(-\tilde{B}_1 t) + \tilde{B}_1 exp(-\tilde{A}_1 t) \).

Assume that the two parameterizations yield the same survival function on this first subinterval, i.e., \( S(t; \tilde{A}_{11}, \tilde{B}_{11}, \frac{\nu_1}{\alpha_1}) = S(t; \tilde{A}_{12}, \tilde{B}_{12}, \frac{\nu_1}{\alpha_2}) \). Then their derivatives
should be identical. That is, for all \( t \in [0, t_1] \)

\[
- \frac{\nu_{11}}{\alpha_{11}} \frac{f_{11}(0, t)}{f_{11}(0, t)} = -\frac{\nu_{12}}{\alpha_{12}} \frac{f_{12}(0, t)}{f_{12}(0, t)},
\]

\[
f_{11}(0, t) = f_1(0, t; \frac{\nu_{11}}{\alpha_{11}}, \tilde{A}_{11}, \tilde{B}_{11})
\]

\[
f_{12}(0, t) = f_1(0, t; \frac{\nu_{12}}{\alpha_{12}}, \tilde{A}_{12}, \tilde{B}_{12})
\]

Plugging in the functions \( f_{11}(.) \) and \( f_{12}(.) \), we get:

\[
\frac{\tilde{A}_{11} \nu_{11}}{\alpha_{11}} \exp((\tilde{B}_{11} - \tilde{A}_{11})t) - 1 \frac{\exp((\tilde{B}_{12} - \tilde{A}_{12})t) - 1} = \frac{\tilde{A}_{12} \nu_{12}}{\alpha_{12}} \exp((\tilde{B}_{12} - \tilde{A}_{12})t) - 1 \frac{\tilde{A}_{12} \nu_{12}}{\alpha_{12}} \exp((\tilde{B}_{12} - \tilde{A}_{12})t) - 1 \frac{\tilde{A}_{12} \nu_{12}}{\alpha_{12}} \exp((\tilde{B}_{12} - \tilde{A}_{12})t) - 1.
\]

If we divide both sides by \( \tilde{A}_{11} \nu_{11} \), then it is obvious this equation has the same form as that in theorem 2.4.2. Here \( c_1 = \tilde{B}_{11} - \tilde{A}_{11}, c_2 = \tilde{B}_{12} - \tilde{A}_{12} \) and \( d_1 = -\tilde{A}_{11}, d_2 = -\tilde{A}_{12} \). As mentioned before, \( \tilde{B}_1 > 0 \) and \( \tilde{A}_1 < 0 \), we have \( c_1 > 0, c_2 > 0, d_1 > 0, d_2 > 0 \). Therefore \( \tilde{B}_{11} - \tilde{A}_{11} = \tilde{B}_{12} - \tilde{A}_{12}, \frac{\tilde{A}_{11}}{\tilde{B}_{11}} = \frac{\tilde{A}_{12}}{\tilde{B}_{12}} = C \neq 1 \), and \( \frac{\tilde{A}_{11}}{\alpha_{11}} = \frac{\tilde{A}_{12}}{\alpha_{12}} \). The first two imply \( \tilde{A}_{11} = \tilde{A}_{12}, \tilde{B}_{11} = \tilde{B}_{12} \). Combining with the third, we conclude \( \frac{\tilde{A}_{11}}{\alpha_{11}} = \frac{\tilde{A}_{12}}{\alpha_{12}} \).

2) Now we move to the second subinterval \([t_1, t_2] \). Let \( t \) change within this interval.

Then equation (2.1) becomes:

\[
\log S(t) = \frac{\nu_1}{\alpha_1} \log \frac{\tilde{B}_1 - \tilde{A}_1}{f_1(0, t_1)} + \frac{\nu_2}{\alpha_2} \log \frac{\tilde{B}_2 - \tilde{A}_2}{f_2(t_1, t)}.
\]

(2.22)

If we carefully examine the right side of the equation, we see that the expression associated with the first subinterval does not involve \( t \). We can treat it as a constant. The only item associated with \( t \) is \( f_2(t_1, t) \). Therefore, we can rewrite the right hand
side of the equation as $-\frac{\omega_2}{\alpha_2} \log f_2(t_1, t) + C_1$. If two survival functions, say $S_1(t)$ and $S_2(t)$, are the same on $[t_1, t_2]$, then $-\frac{\omega_1}{\alpha_1} \log f_{21}(t_1, t) + C_{11} = -\frac{\omega_2}{\alpha_2} \log f_{22}(t_1, t) + C_{12}$. The definitions for $f_{21}, f_{22}$ are similar to those of $f_{11}, f_{12}$. Taking derivatives on both sides, we obtain

$$\frac{f_{21}'(t_1, t)}{f_{21}(t_1, t)} = \lambda_1 \frac{f_{22}'(t_1, t)}{f_{22}(t_1, t)},$$

where $\lambda_1 = \frac{\omega_2}{\omega_2} \frac{\alpha_1}{\alpha_2}$.

Recalling the definition of $f_2$, we obtain

$$\frac{f_{21}'(t_1, t)}{f_{21}(t_1, t)} = \frac{-\tilde{A}_{21}(\exp((\tilde{B}_{21} - \tilde{A}_{21})(t - t_1)) - 1)}{\exp((\tilde{B}_{21} - \tilde{A}_{21})(t - t_1)) - \frac{d_1}{\tilde{B}_{21}}},$$

$$\frac{f_{22}'(t_1, t)}{f_{22}(t_1, t)} = \frac{-\tilde{A}_{22}(\exp((\tilde{B}_{22} - \tilde{A}_{22})(t - t_1)) - 1)}{\exp((\tilde{B}_{22} - \tilde{A}_{22})(t - t_1)) - \frac{d_2}{\tilde{B}_{22}}}.$$

According to theorem 2.4.2, we can show that $\tilde{A}_{21} = \tilde{A}_{22}, \tilde{B}_{21} = \tilde{B}_{22}$ and $\frac{\omega_1}{\alpha_1} = \frac{\omega_2}{\alpha_2}$. Here $c_1 = \tilde{B}_{21} - \tilde{A}_{21} > 0$. Similarly, $c_2 = \tilde{B}_{22} - \tilde{A}_{22} > 0$. $d_1 = -\frac{d_1}{\tilde{B}_{21}} > 0$ and $d_2 = -\frac{d_2}{\tilde{B}_{22}} > 0$. Finally $\lambda = \frac{\omega_2}{\omega_2} \frac{\alpha_1}{\alpha_2}$. The equalities $c_1 = c_2$ and $d_1 = d_2$ give $\tilde{A}_{21} = \tilde{A}_{22}$ and $\tilde{B}_{21} = \tilde{B}_{22}$. Combining with $\lambda = 1$, we have $\frac{\omega_1}{\alpha_1} = \frac{\omega_2}{\alpha_2}$.

3) When $t \in [0, t_1]$, the equality $S_1(t) = S_2(t)$ suggests that $\frac{\omega_1}{\alpha_1}, \tilde{A}_1$, and $\tilde{B}_1$ are uniquely determined by the survival function. We showed that when $t \in [t_1, t_2]$, the other three parameters are identifiable, namely, $\frac{\omega_2}{\alpha_2}, \tilde{A}_2$ and $\tilde{B}_2$. So is $f_2()$. In order to make the equality (2.22) hold for any $t \in [0, t_2]$, one more requirement must be met: $f_{11}(0, t_1) = f_{12}(0, t_1)$. Recall from the definition of $f_i(0, t_1)$, we have
\[ f_1(0, t_1) = -\tilde{A}_1 e^{\tilde{B}_1(0-t_1)} + \tilde{B}_1 e^{\tilde{A}_1(0-t_1)} + y_1(e^{\tilde{B}_2(0-t_1)} - e^{\tilde{A}_1(0-t_1)}). \]

The equality between \( f_{11} \) and \( f_{12} \) implies that \( \tilde{y}_{11} = \tilde{y}_{12} \) if the two sets of parameters are to produce identical survival functions. However, \( y_1 = \frac{\alpha_1}{\alpha_2} \partial_u \log f_2(u, t)|_{u=t_1}. \)

The function \( f_2(t_1, t) \) is unique once the two survival functions agree on \([0, t_2]\). Hence the two ratios must be equal, \( \frac{\alpha_{11}}{\alpha_{21}} = \frac{\alpha_{12}}{\alpha_{22}}. \)

4) Proceeding in the same fashion and using theorem 2.4.2, we prove that in each subinterval there are three identifiable parameters. This sums up to \( 3K \). The remaining \( K - 1 \) identifiable parameters are \( \frac{\alpha_{i-1}}{\alpha_i}, i = 2, \cdots, K. \)

In summary, \( \tilde{A}_i, \tilde{B}_i, \frac{\mu_i}{\alpha_i}, \) and \( \frac{\alpha_{i-1}}{\alpha_i} \) are identifiable given the two survival functions are exactly the same on the positive real line. \( \diamond \)

When clonal expansion is introduced to the multi-stage stochastic model, the complexity of mathematical derivation increases tremendously. Therefore, most of the nice theoretical results on multistage models are limited to the two-stage clonal expansion model.

As some experimental data have indicated, there are more than two stages involved in carcinogenesis [18]. In the final section of this chapter, we will cover some of the research results in multistage \((n \geq 3)\) clonal expansion models and discuss difficulties encountered when applying these models to human cancer incidence data.
2.5 Multistage clonal expansion models

The two-stage clonal expansion model has been extended to the multi-stage clonal expansion models. Tan, Moolgavkar, Luebeck, and Little et al. ([16], [29] and [17]) are the main contributors to the development of theories for multi-stage models. All derivations are still based on PDEs for pgf. Because of the nice mathematical properties of Markov processes, most of the derivations and arguments set up in the TSCE model can be followed easily. However, there are no closed-form solutions for the survival function of the waiting time for a tumor to arise. Moreover, the non-identifiability problem gets worse if each stage allows independent parameters and the fitted data is age-specific incidence rate. Due to ethical issues, there are no experimental data on human cancer recording the dynamic change of intermediate cells before cancer occurs. Therefore, for human cancer data, two-stage models are still most frequently used.
Chapter 3

Alternative carcinogenesis models

In the previous chapter, we surveyed the one-path carcinogenesis model. However, some experimental data and carcinogenesis theories lean towards multi-pathway models. For instance, lung cancer has two groups: small cell and non-small cell. “Each grows and spreads in different ways and is treated differently” [19]. Also, leukemia can be divided into two types: acute and chronic. Within each type it can be further classified as either lymphocytic or myelogeneous leukemia. The differences in symptoms and treatments among different kinds of cancer imply that cancerous cells can progress through different paths and end at different destinations during carcinogenesis. Further discussion about multiple pathways can be found in Tan [29] and Mao [18]. In this chapter, we focus on the derivation of a multi-pathway stochastic model in the framework of TSCE.
3.1 Three-stage two-path stochastic model

We start by considering a simple three-stage two-pathway carcinogenesis model. Assume there are three stages of abnormal cells: type 1, 2 and 3. The higher the level is, the more malignant the cells are. The cells can jump either from 1 to 2 and then from 2 to 3 or from 1 to 3 directly (illustrated in figure 3.1). Once cells are in stage 3, they will stay there forever. For reasons that will be clear later, we will consider the simpler problem here: Our main interest will be the distribution of the waiting time $T_c$ until the first type 3 cell arises starting with a single type 1 cell. The cells have cell division rate $\lambda$ per year and cell type change rate $p_0$ per year. In this chapter, all biological parameters are assumed to be constant.

Figure 3.1: Three-stage transition paradigm

3.2 Cumulative density function (CDF) of $T_c$

In order to derive the CDF for $T_c$, we need to track all possible events which could lead to the first type 3 cell. There are only two possibilities: The first type 3 cell arises from either a type 2 cell or a type 1 cell. For clarity, when we say a type 1 cell
gives rise to a type 3 cell, we mean that this type 1 cell or its descendants give birth to a new type 3 cell.

Let us assume that there is a single type 1 cell at time 0 in the three compartment system. This cell may proliferate or jump to another stage. Suppose that the first jump happens at time $T'$. Then we need to consider two possible destinations for this jump in order to derive the formula for $T_c$. This "mutated" cell can fall either into the type 2 pool with probability $p$ or the type 3 pool with probability $1 - p$. If it goes directly to stage 3, then a malignant cell arises. We assume this malignant cell will develop into a tumor in a short time period with probability one. So the waiting time $T_c$ is $T'$. The distribution of $T'$ has been extensively studied in the TSCE model and a closed-form solution has been given for the constant parameter setting (theorem 2.4.1 in Chapter 2). If this "mutated" type 1 cell falls in the type 2 cell pool, the waiting time $T_c$ depends not only on $T'$ but also on the number of type 1 cells present at time $T'$. The regeneration process of the ancestor type 1 cell is assumed to follow a birth and death process. After time $T'$, the ancestor type 1 cell has already built a clone of its own. According to the stochastic nature of a birth and death process, the number of type 1 cells born by time $T'$ is a random variable (r.v.) depending on $T'$ and the parameters of the birth and death process. Let $X(T')$ be the number of type 1 cells at time $T'$. In addition, there exists one type 2 cell. We need to wait for one of these type 1 cells or this new type 2 cell to give rise to a type 3 cell. If this extra waiting time is denoted $Y(X(T'))$, then $T_c = T' + Y(X(T'))$. Using probability
notation, we can rewrite the waiting time $T_c$ as:

$$T_c = I_{1-p}T' + I_p(T' + Y(X(T'))),$$  \hspace{1cm} (3.1)$$

where $p$ is the probability that the jump destination is the type 2 cell pool given there is a jump.

As in the TSCE model, we continue to use the assumption that the waiting time for one cell to divide or jump is exponentially distributed. Let $T_{23}'$ be the waiting time for the first type 3 cell to arise from a single type 2 cell and let $T_i$ be the $i$-th iid copy of $T_c$. It is obvious that $T_i \sim T_c$ because every type 1 cell is identical. Using the Markov property of this multi-path and multi-stage model, we can derive the formula for the random variable $T_c$ as follows.

**Theorem 3.2.1** In the homogeneous three-stage two-pathway model (in which all parameters associated with the transition and birth-death processes are constant) the cumulative density function (CDF) of $T_c$ is given by:

$$F_{T_c}(t) = (1 - p)F_{T'}(t) + p \int_0^t f_{T'}(s)(1 - (1 - F_{T_{23}'}(t - s))g(t, s; T_c))ds,$$

where $g(t, s; T_c) = \frac{e^{-(a-\beta)z}(z-\beta)}{1+(z-\beta)(e^{-\alpha z}\frac{\alpha}{a-\beta})}$, $z = (1 - F_{T_c}(t - s))$, $\alpha, \beta$ are the birth and death rates associated with a type 1 cell, $p$ is the probability that a type 1 cell will jump to the type 2 cell pool given there is a jump, $T'$ is the waiting time for a single type 1 cell to give rise to a new type of cell, and $T_{23}'$ is the waiting time for a single type 2 cell to give rise to a type 3 cell.
Proof: According to the equation 3.1, the CDF of $T_c$ can be decomposed into two parts:

\[
P(T_c \leq t) = (1 - p)P(T' \leq t) + pP(T' + Y \leq t)
\]
\[
= (1 - p)P(T' \leq t) + p \int_0^\infty f_{T'}(s) P(s + Y \leq t)ds
\]
\[
= (1 - p)P(T' \leq t) + p \int_0^\infty f_{T'}(s) P(Y \leq t - s)ds
\]

Let us consider the distribution of $Y$:

\[
P(Y \leq t - s)
\]
\[
= \sum_{n=0}^\infty \{P(\min(T'_2, \min_{0 \leq i \leq n} T_i) \leq t - s)P(X(s) = n)\}
\]
\[
= 1 - \sum_{n=0}^\infty \{P(\min(T'_{23}, \min_{0 \leq i \leq n} T_i) > t - s)P(X(s) = n)\},
\]
\[
= 1 - P(T'_{23} > t - s) \sum_{n=0}^\infty \{P(\min_{0 \leq i \leq n} T_i > t - s)P(X(s) = n)\}, \text{ because } T'_{23} \text{ is independent of } T_i \text{ and } X(s).
\]

Further,

\[
P(Y \leq t - s) = 1 - (1 - F_{T'_{23}}(t - s)) \sum_{n=0}^\infty (\prod_{i=1}^n [1 - F_{T_i}(t - s)])P(X(s) = n)
\]
\[
= 1 - (1 - F_{T'_{23}}(t - s)) \sum_{n=0}^\infty P(X(s) = n)(1 - F_{T_c}(t - s))^n \text{ because } T_i \sim T_c.
\]

Finally, \(P(Y \leq t - s) = 1 - (1 - F_{T'_{23}}(t - s))g(t, s; T_c),\)

where

\[
g(t, s; T_c) = \sum_{n=0}^\infty P(X(s) = n)(1 - F_{T_c}(t - s))^n.
\]

For fixed $t$ and $s$, $0 \leq 1 - F_{T_c}(t - s) \leq 1$. If we replace $1 - F_{T_c}(t - s)$ with $z$, then it is clear that $g(t, s; T_c)$ has the form of the pgf of $X(s)$. For a stochastic process
with birth rate $\alpha$ and death rate $\beta$, the pgf for the number of particles in the system at time $t$ is:

$$G(z, t) = \frac{e^{-(\alpha-\beta)t}(z - \frac{\beta}{\alpha})}{1 + (z - \frac{\beta}{\alpha})(e^{-(\alpha-\beta)t} - \frac{\alpha}{\alpha-\beta})},$$

where $z \in (0, 1)$ and the number of particles at time 0 is 1. Thus the function $g(.)$ has the analytical form:

$$g(t, s; T_c) = \frac{e^{-(\alpha-\beta)s}(z - \frac{\beta}{\alpha})}{1 + (z - \frac{\beta}{\alpha})(e^{-(\alpha-\beta)s} - \frac{\alpha}{\alpha-\beta})},$$

where $z = 1 - F_{T_c}(t - s)$.

Clearly, $g(t, t; T_c) = 1$ and $g(t, 0; T_c) = 1 - F_{T_c}(t) = z$ since $P(X(0) = 1) = 1$.

The random variable $T_{23}$ is similar to $T'$ but may differ in parameters (transition rate, birth and death rate). Plugging the CDF of $Y$ into the equation for $P(T_c \leq t)$, we have

$$F_{T_c}(t) = (1 - p)F_{T'}(t) + p \int_0^t f_{T'}(s)(1 - (1 - F_{T_{23}}(t - s))g(t, s; T_c))ds.$$ 

The analytical forms for $F_{T'}, f_{T'}$ and $F_{T_{23}}$ can be found in Chapter 2. ♦

### 3.3 Numerical computation of the CDF of $T$

We have an equation for the CDF of $T_c$ in theorem 3.2.1, but the right hand side involves an unknown value of $F_{T_c}(t - s)$ in the pgf part. To find the solution for $F_{T_c}(.)$ we need a numerical algorithm to compute it recursively. Here we assume $\beta = 0$. A similar derivation can go through for the case when $\beta \neq 0$. 
In general, for any point \( n \Delta t \), partition the whole integral interval \([0, n \Delta t]\) into \( n \) small subintervals, \( t_0 = 0, t_i - t_{i-1} = \Delta t \) and \( t_n = n \Delta t \), then the recipe for computing \( F_{T_c}(t_n) \) is given by:

i) For \( i = 1 \), \( F_{T_c}(t) \) is approximated by:

\[
F_{T_c}(\Delta t) \approx (1 - p) F_{T'}(\Delta t) + p F_{T'}(\Delta t/2)(1 - (1 - F_{T_{23}}(\Delta t/2)) G(\Delta t, \Delta t/2, T_c)) \Delta t,
\]

where \( G(\Delta t, \Delta t/2, T_c) = \frac{e^{-\alpha \Delta t/2} z}{1 + z(e^{-\alpha \Delta t/2} - 1)}, z = 1 - F_{T_c}(\Delta t/2) \)

and

\[
F_{T_c}(\Delta t/2) \approx (1 - p) F_{T'}(\Delta t/2) + p f_{T'}(\Delta t/2)(1 - (1 - F_{T_{23}}(\Delta t/2))) \Delta t/2.
\]

By the properties of CDFs, \( F_{T_c}(0) = 0 \). Hence \( F_{T_c}(s) \approx 0 \) when \( s \leq \Delta t/2 \).

ii) For \( i = 2, \cdots, n \),

a) compute \( z = 1 - F_{T_c}(t_i - t_j) \) and \( g(t_i, t_j; T_c) \), for \( j = 1, \cdots, i - 1 \);

b) compute \( h(t_j, t_i) \), for \( j = 1, \cdots, i \),

where \( h(t_j, t_i) = f_{T'}(t_j)(1 - (1 - F_{T'}(t_i - t_j)) g(t_i, t_j, T_c)) \);

c) compute

\[
F_{T_c}(t_i) = \{(1 - p) F_{T'}(t_i) + p \sum_{j=1}^{i-1} (h(t_j, t_i) + h(t_{j+1}, t_i)) \Delta t/2
\]

\[+ph(t_1, t_i) \Delta t/2 + p \Delta t f_{T'}(0)/2 F_{T_{23}}(t_i)\}/y_i,
\]

where \( y_i = 1 - p \Delta t f_{T'}(0)(1 - F_{T_{23}}(t_i))/2 \).

End of \( i \) loop.
We know how to calculate the CDF of $T_c$. However, only the random variable $T$, the waiting time from birth to the rise of the first malignant cell, can be observed in real life. The assumption that the transition from the type 0 cell pool to the type 1 cell pool follows a Poisson process with density $\nu(t)$ enables us to build a bridge to connect the two random variables $T_c$ and $T$ by using the filtered Poisson process property through this formula:

$$S_T(t) = \exp\left(-\int_0^t \nu(s) F_{T_c}(t-s) ds\right).$$

For more rigorous derivation of the formula, please see Parzen [23] and Hanin [7]. After some algebra, we can show that the hazard function of $T$ can be computed from

$$\lambda(t) = \int_0^t \nu(s) f_{T_c}(t-s) ds.$$

In the end, we will check the convergence of our numerical solution derived above. We will set two different step sizes, $\Delta t = 0.01$ and $0.005$, compute the hazard functions of $T$ at given age points, and compare the numerical values obtained from the two step sizes. Visually, it seems that this primitive integration algorithm converges when the step size is moderately small, i.e. $\Delta t = 0.01$ (in years) (figures 3.2 and 3.3).
Figure 3.2: Comparison of the hazard rate with different step sizes using Deng’s numerical algorithm (Scenario I). In the plot, alpha is the same as p in the theorem. The parameter $p_0$ is the probability for a single cell to jump in one year, growb is the net growth rate (same as $\alpha - \beta$ in the theorem) for the intermediate cells and $v$ is the first transition rate (from normal cell to intermediate cell).

### 3.4 Tan’s multi-pathway model and approximation formula

An active researcher in carcinogenesis modeling, Tan [29] proposed the idea of a multi-pathway clonal expansion model a decade ago. He established a similar pathway with more compartments than the three-stage two-path model (figure 3.4). If one ignores $I2$ and sets $\gamma(t) = 0$, then his model is essentially the same as my three-stage two-path
Figure 3.3: Comparison of the hazard rate with different step sizes using Deng's numerical algorithm (Scenario II). In the plot, alpha is the same as $p$ in the theorem. The parameter $p_0$ is the probability for a single cell to jump in one year, growb is the net growth rate (same as $\alpha - \beta$ in the theorem) for the intermediate cells and $\nu$ is the first transition rate (from normal cell to intermediate cell).

Tan's model is more complicated and involves more parameters. In his book, Tan took the traditional pgf approach to solve for the survival and hazard functions of the waiting time $T$ for the rise of the first malignant cell. First he set up the joint pgf for the numbers of type 0, type 1, type 2, and type 3 cells, then made use of the Markovian properties of these processes and derived differential equations to write the distribution of $T$ as a complex function of some auxillary functions, which only can be solved numerically from a PDE system. Even in the simplest case, when all
parameters are constant, the PDE is reduced to a Ricatti equation. Tan did not give any closed-form solution for the hazard function of $T$. He pointed out that solving those differential equation(s) is not trivial and that a numerical method is necessary. He also gave the following approximation formula:

$$
\lambda(t) = \int_0^t \nu(u)[\beta_1(t)m_1(u, t) + \beta_3(t)m_2(u, t)]du,
$$

$$
m_1(u, t) = \exp(\int_u^t (\alpha(x) - \beta(x))dx),
$$

$$
m_2(u, t) = \int_u^t w_1(y)\exp\{\int_u^y (\alpha_1(x) - \beta_1(x))dx + \int_y^t (\alpha_2(x) - \beta_2(x))dx\}dy,
$$

where $\beta_1(t)$, $\beta_3(t)$ and $w_1(t)$ are the transition rates from type 1 cell to type 3 cell, from type 2 cell to type 3 cell, and from type 1 cell to type 2 cell respectively. The parameters $\alpha_1$, $\beta_1$, $\alpha_2$, and $\beta_2$ are the birth and death rates for type 1 and type 2 cells respectively (see the illustration in figure 3.4).

Tan claimed that the approximation will be good if the net proliferation rate $\alpha - \beta$ is no greater than 0.1 and the mutation (transition) rate for a single cell is between $10^{-8} \sim 10^{-6}$. Therefore, under his suggested conditions, I created several special parameter sets, computed hazard rates using the two approaches: Deng's numerical algorithm and Tan's approximation and compared the numerical results. When age is not very high, the two give almost identical results (figure 3.5 and 3.6). However, the discrepancy occurs when age passes a certain limit (figure 3.7). It is worth mentioning that when $w_1(t)$ or $\beta_1(t)$ is zero, the three-stage two-path model is reduced to the TSCE model. A comparison between Tan's approximation and
the closed-form solution for the reduced TSCE model also shows a similar pattern: Tans formula overestimates the hazard rates as age increases. It is hard to conclude which method, Deng’s numerical algorithm or Tan’s approximation, is better or more accurate when no other reference approach is available. But from figure 3.8, we are more confident in Deng’s algorithm. The interpretation of parameters in my model is not the same as those in Tan’s model. After some conversion, it can be shown that the two models are equivalent.
Figure 3.5: Comparison of the hazard rates using Tan’s approximation and Deng’s algorithm

3.5 Fitting the three-stage two-path model to the Texas lung cancer data

The three-stage two-path model can also fit cancer incidence data well. After searching for the best fit through a fine grid, we found the fit of this model to be as good as that of the TSCE model (figure 3.9 and 3.10). As mentioned at the beginning of this chapter, three or more stages can occur in carcinogenesis. Moreover, multiple paths are more reasonable explanation for different types of lung cancer. Thus, in those circumstances, the formula developed for multi-stage and multi-path models is useful for the study of carcinogenesis.
Figure 3.6: Comparison of the hazard rates using Tan’s approximation and Deng’s algorithm.

Figure 3.7: Comparison of the hazard rates using Tan’s approximation and Deng’s algorithm.
Figure 3.8: Comparison of the hazard rates for TSCE using Tan's approximation and the closed-form formula.

Figure 3.9: Nonlinear Regression Estimation (TSCE model) for female whites 1996-2000
Figure 3.10: Nonlinear Regression Estimation (3-stage 2-path) for female whites 1996-2000
Chapter 4

Evaluating the effects of smoking and genetic susceptibility

In this chapter we will evaluate the impacts of cigarette smoking and genetic susceptibility on lung cancer incidence through a two-stage clonal expansion model with a piecewise constant biological parameter setting. The assessment of the impact of each factor on the age-specific incidence of lung cancer will allow us to predict the probability of an individual to have lung cancer at a certain age provided his/her smoking history and genetic susceptibility levels. Also from the population point of view, we are able to forecast the incidence rate of lung cancer in future years in a population of interest if reliable knowledge on the distribution of each covariate is available. Several research groups have applied the framework of TSCE to real data and combined it with estimation approaches to evaluate the impact of cigarette smoking, radon expo-
sure, or other environmental risk factors on the probability of developing lung cancer at certain age ([9], [12] and [16]). However, genetic factors could also play a role in the carcinogenesis of lung cancer. The capability to metabolize chemical compounds from tobacco smoke and the ability to repair damage caused by environmental exposure can vary from individual to individual. These capacities are mainly genetically determined. Our new contribution is to take genetic susceptibility into account. We study how genetic factors modify the risk that cigarette smoking confers and what role genetic factors play in transition rates in the TSCE model for never smokers. Most early work applying TSCE to real data merely demonstrated the validity of the estimation approach under a simulation paradigm and did not show how sampling errors and right censoring affect the precision of estimates. In this section, we will explore some aspects of these issues.

4.1 Summary of prospective data and case-control data

Most cancer registries only keep records of patients’ age, race, and gender. Few large datasets supply information on an individual’s smoking history. CPS-II study data (1982-1988) is one of the most recently published datasets. It gives age-specific lung cancer incidence data for never and current smokers, enabling us to estimate parameters of the TSCE model for populations with similar smoking status. Such
estimates should be more plausible and meaningful. CPS-II data is also prospective, and the raw data contains a huge amount of individual records on age, race, gender, smoking history, cancer history, etc. The figures we are interested in are the summarized population-based death rates for lung cancer. Besides smoking history, genetics are another risk factor we will evaluate. However, datasets with genetic susceptibility measurements are hard to obtain since the time, lab techniques, and costs of quantifying genetic susceptibility for a population are tremendous. Through collaboration, we gained access to a dataset based on a case-control study conducted in the M.D. Anderson Cancer Center Epidemiology Department. The case-control data includes detailed individual information such as smoking history, environmental exposure, DNA Repair Capacity (DRC), age, etc. All subjects in the two datasets are white. Therefore all analyses and inferences will be for the white population only. In the next section, we give more detailed descriptions of the data sources.

4.1.1 Cancer Prevention Study II data

Great achievements in medicine and improvements in public hygiene greatly helped the control of several deadly diseases prevalent in the early twentieth century. In the early 1950s, cancer emerged as a new hazard to public health and a new problem for physicians and scientists. The American Cancer Society carried out two largest prospective cancer prevention studies. The primary goals of these studies were to identify the main risk factors for various types of cancers, to capture the time pattern
of cancer mortality rates, and to study the change in the distribution of prevalent cancers. The first study, Cancer Prevention Study I was between 1959 and 1965. The second one, Cancer Prevention Study II, was the largest and most recent prospective study of smoking and related diseases. There were 1,185,106 subjects in this study. The subjects were recruited nationwide by ACS volunteers, starting in 1982. Each subject was over age 30 with at least one family member older than 45. Usually once one family member was enrolled in the program, others in the household meeting the age requirement were asked to join as well. The subjects were followed for six years through volunteer contact every two years, from the month of enrollment in 1982 through August 31, 1988. In the entire cohort, 79,802 participants (6.7 percent) died, 1,083,600 (91.4 percent) survived, and 21,704 (1.8 percent) were lost to follow-up during this interval [31]. The data we have access to include analyses restricted to never smokers and current cigarette smokers. In the final analysis results, they excluded former smokers because of the large heterogeneity among former smokers with respect to initiation and cessation ages. To construct homogeneous former smoker groups, the number of subjects falling into each category would not be sufficient to get stable estimates. The data to be used are taken from tables in Appendix 5, 6 of Chapter 4 [31]. These tables include stratified death rates of lung cancer, for the whites only, excluding prevalent cancers.
4.1.2 M.D. Anderson case-control study

A case-control study is being carried out in the Epidemiology Department of M.D. Anderson Cancer Center by Dr. Spitz, Dr. Wu, and Dr. Wei in order to study lung cancer and its association with the genetic composition of patients. The measurements of DNA repair capacity (DRC) were achieved through the Host Cell Reaction Assay. These bio-assays were specially designed to quantify DRC by counting DNA adducts after exposing cells to chemical compounds for a given period of time. Briefly, a blood sample was taken, “the frozen lymphocytes of each patient are thawed and processed to ensure a cellular viability of > 80%. The cells are then stimulated so that they have taken up the plasmids. The number of viable, large lymphoblasts in the culture for each sample is counted to calculate the blastogenic rate. Duplicate transfections with either untreated plasmids or BPDE-treated plasmids are always performed ... DRC is reported as the ratio of the radioactivity of cells transfected with BPDE-treated plasmids to that of cells transfected with untreated plasmids.”

As to the details of the biological aspects of the designs and experiments, please see Dr. Spitz’s paper [28]. The M.D. Anderson dataset contains a total of 1200 records of case-control subjects from the white population. Two hundred seventy two female patients were one-to-one matched with 272 female controls on age and smoking status. Similarly, 328 male patients were matched with 328 male controls. The age match was not exact but up to the same age group (five year bin). Due to practical difficulties, precise smoking intensity and smoking duration matching were not enforced. Among
all patients, about 50% are former smokers and around 30% are current smokers. The rest were never smokers and recent quitters. The dataset provides detailed records of smoking initiation, cessation age, number of cigarettes per day, actual smoking years, last time of quitting, the reason for quitting, the method of quitting, environmental exposure to risk factors other than cigarette smoking, family cancer history, and so on. There are some minor to substantial missing values for the measurements of DRC. There is more than 25% missing among controls of current smokers. The missing is more common for controls than cases. The percentages of missing DRC range from 7% to 30%. The age distribution among lung cancer patients is not typical for the white population. The case-control data has a higher percentage of young patients than would be suggested by age-specific lung cancer incidence rates for Texans.

### 4.1.3 Exclusion of former smokers’ data

We discarded most of the variates in the case-control study data. Our original plan included the study of quitting effects on the risk of developing lung cancer for former smokers. Although half of the case-control subjects are former smokers, the construction of age-specific incidence rates from this case-control data is not feasible. CPS-II does not provide this information either. The only remaining possibility is to recover the incidence rates for former smokers from the combined incidence data on never, former, and current smokers. Although we have the combined Texas lung cancer incidence data (1996-2000), compatibility is our first concern. It is clear that there are
great disparities between the CPS-II data and the Texas lung cancer incidence data (see figure 4.1). The mixed incidence rate in the Texas data is as high as the mortality for current smokers alone in CPS-II. In this thesis, we do not distinguish between the mortality rate and the incidence rate since lung cancer has very poor prognosis. One hypothesis is that the large gap between the Texas data and the CPS-II data is caused by geographic differences since the CPS-II data covers almost all states in the U.S., while the Texas data includes only one state. Other data sources such as the SEER lung cancer incidence data, collected from nine registries, may serve as a better representation of lung cancer incidence for the U.S. white population. However, we noticed that the SEER incidence data from 1996-2000 also provides a different image of lung cancer incidence compared to that in CPS-II (1982-1988) (figure 4.1).

Hence, the major difference more likely stems from the time change than from the population shift. Incompatibility of the Texas and the CPS-II data forces us to abandon the idea of making inference on lung cancer incidence in former smokers since the TSCE model requires incidence data based on a population with similar smoking status. From now on, we focus on the statistics based on never and current smokers in this case-control dataset.

Before we begin the analysis of real data, we first lay down the overall estimation strategy for the parameters associated with genetic factors and cigarette smoking. Then, using simulation, we investigate estimation properties when applying incidence data alone and considering right censoring and sampling errors.
Figure 4.1: SEER, CPS-II and Texas Lung Cancer data comparison. Never and current smokers mortality comes from CPS-II. The former smokers’ incidence rate is computed by subtracting never and current smokers mortality rates from the combined Texas incidence rates.
4.2 Overall estimation strategy

First, we set up a sample response function between all covariates and biological parameters and use it to illustrate our estimation strategy:

\[
\mu(t) = \mu_0[(1 + a_2I(SM; t))(1 + a_1(1 - drc))(1 - I(SM; t)) + (1 + a_4(1 - drc))I(SM; t)];
\]

\[
\nu(t) = \nu_0[(1 + a_3I(SM; t))(1 + a_1(1 - drc))(1 - I(SM; t)) + (1 + a_4(1 - drc))I(SM; t));
\]

\[
\alpha(t) = \alpha_0(1 + a_5 \ast (\log(SD) - a_6)I(SM; t));
\]

\[
\beta(t) = \beta_0(1 + a_5 \ast (\log(SD) - a_6)I(SM; t));
\]

SD is the smoking density in number of cigarettes per day and drc is the DNA repair capacity level. SM is the smoking status and \(I(\cdot)\) is an indicator function. That is, \(I(SM; t) = 1\) in the presence of cigarette smoking at time \(t\) and \(I(SM; t) = 0\) in the absence of smoking.

Analysis for each gender will be conducted separately. The coefficients \(a_2, a_3, a_4, a_5,\) and \(a_6\) only relate to cigarette smoking. Hence, they can only be estimated from current smokers’ data.

We start from the simple case of never smokers. Using never smokers’ data, we estimate not only three of the biological parameters \(\nu_0, \mu_0,\) and \(\alpha_0\) with \(\beta_0\) given but also the coefficient \(a_1\) corresponding to genetic factors. Never smokers’ data yields estimates for the baseline parameters and for the coefficient of DRC. Plugging the estimates into the response function, we run the regression for current smokers and estimate the coefficients of the risk factors. There is variation in initiation age and
smoking intensity among current smokers. Some distributions of smoking intensity and initiation were assumed by considering incomplete survey data from the CPS-II report. The distribution of DRC among current smokers is estimated from that of current smokers in the case-control study. Combining this with stratified smokers’ incidence figures and overall age-specific lung cancer incidence rates for current smokers, we test the association between cigarette smoking and enhanced proliferation and transition rates.

Due to nonidentifiability, we must fix one of the parameters. We choose the death rate $\beta_0$. This parameter is selected to make the first transition rate, $\nu_0$, consistent with results reported in one of the papers on lung cancer carcinogenesis [12].

### 4.3 Random noise and the trade-off between $\nu$ and $\mu$

In this section, we begin with the constant parameter TSCE model without considering any covariates in the response function, introduce the technical set-up, and review problems caused by right censoring and sampling errors. Although the TSCE model has been applied many times to assess risk factors in various circumstances, most of the papers on the application of the TSCE models avoided to address explicitly how the statistical issues were handled and to explain some shortcuts in data analysis. We feel it is necessary to learn more about the precision of estimates and the sources of
variation in simulation settings. Before we can explore the identifiability of parameters and the errors in estimation caused by random noise in the incidence rates, we must introduce the basic tools: the optimization functions, the objective function, the estimation approach, and the response function.

4.3.1 Two optimization functions and the logit transformation for parameters

The two Matlab optimization functions utilized are: \textit{fminsearch} and \textit{fminunc}. The function \textit{fminsearch} uses the Nelder-Mead simplex search algorithm, while the function \textit{fminunc} provides different methods for choosing the search direction in the Quasi-Newton algorithm. In our case, a line-search algorithm is implemented when the Hessian matrix is not provided. Both optimization functions return a local optimal value and allow unrestricted parameter values. Since the model was built for biological processes, there must be boundaries set for all biological parameters and coefficients. That is, each parameter is only allowed to vary within a closed interval, say \([a, b]\). To resolve the conflict between the optimization function settings and the biological restrictions, a transformation, such as the logit function, is needed to map a parameter on a closed interval \([a, b]\) to a value on \((-\infty, \infty)\).
4.3.2 Least Squares Estimation (LSE) and the objective function

We will use Least Squares Estimation (LSE) as our main estimation approach since we do not see individual information from the CPS-II data. For clarity, let us define our objective function:

\[ f_1 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i, \]

where \( O_i \) is the observed number of lung cancer cases in the \( i \)th age group and the \( E_i \)'s are the predicted numbers of lung cancer cases based on the TSCE model given certain parameter combinations and the population size under study. Let the population size be \( 10^7 \) unless otherwise stated.

4.3.3 Response function

When there is no smoking factor in the analysis, the response function depending on DRC, smoking intensity, and the four biological parameters is reduced to:

\[ \mu(t) = \mu_0(1 + a_1(1 - drc)); \]

\[ \nu(t) = \nu_0(1 + a_1(1 - drc)); \]

\[ \alpha(t) = \alpha_0; \]

\[ \beta(t) = \beta_0. \]

Returning to the non-identifiability issue, in the constant parameter TSCE model with known death rate, the other three parameters can be identified from the inci-
idence data. In the following sections, we show that this conclusion might not be valid when the incidence data is not perfect. In fact, every single point in incidence data represents not the hazard value at that point, but the average probability of developing cancer within a five-year bin. Moreover, incidence data is always contaminated. We must examine two types of incidence data in the simulation studies: 1) ideal incidence rate data without random noise and 2) incidence rate data with random noise due to sampling errors.

4.3.4 Estimates of $\mu$ and $\nu$ in the absence of random noise

When the observed incidence is the true incidence, the two optimization functions seem to adequately pinpoint the right values. But the optimization function fminsearch does better finding the true values of $\mu$ and $\nu$ than fminunc in the absence of random noise in the incidence rates.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$\nu$</th>
<th>$\mu$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>fminsearch</td>
<td>0.3</td>
<td>1e-7</td>
<td>0.36</td>
<td>0.28</td>
</tr>
<tr>
<td>fminunc</td>
<td>0.3344</td>
<td>8.9733e-8</td>
<td>0.36</td>
<td>0.28</td>
</tr>
<tr>
<td>True parameter value</td>
<td>0.3</td>
<td>1e-7</td>
<td>0.36</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Table 4.1: Least Squares Estimation: Comparison of Two Optimization Functions (20 runs)

The function fminsearch runs slower than fminunc. The relatively poorer per-
formance of \textit{fminunc} might result from the flatness of the objective function at the optimal point or from multiple minimum points. A wide range of parameter combinations can yield very similar values of the objective function. This property of the objective function makes it harder to estimate the true parameter values. We can use \textit{fminunc} first to roughly search for a good initial seed and then plug this seed into \textit{fminsearch}. By doing so, we can speed up the search process without sacrificing much precision.

4.3.5 Trade-off between parameters $\mu$ and $\nu$ in the presence of random noise

We now switch our focus from ideal incidence data to data with sampling errors. In the TSCE model, the net growth rate $\alpha - \beta$ is the slope of the incidence curve in logarithmic scale and thus is easy to estimate. The other two identifiable parameters are $M$ and $B$. The parameter $M$ is determined by $\nu_0\mu_0$, and $B$ is partially related to the asymptotic height of the incidence curve. Human cancer incidence rates do not always display a plateau in the tail of the curve because of a shortage of information on old people (more than 85 years old). Hence, the parameter $B$ might not be inferred correctly from partially observed data. Heidenreich also pointed out in his 1996 paper [10]: If the incidence curve has not reached a plateau, only two parameters might be identifiable. How the completeness of the data affects parameter identifiability and precision of estimates was not discussed in his paper. We design two scenarios: 1)
observed incidence rates up to 85, a realistic setting, and 2) a full observed incidence curve recording survival time up to the non-realistic age of 220. The incidence curve in the second scenario would certainly have reached its asymptotic height. Then we add Poisson-distributed noise to the incidence counts assuming that the population size is $10^7$. The optimization procedures \texttt{fminunc} and \texttt{fminsearch} are alternatively used to find LSE for $\mu$ and $\nu$. Both optimization functions rely on the initial input and can only find a local minimum. Twenty random seeds are generated as initial inputs and the entire search procedure is repeated twenty times to improve the chance of locating a global minimum. The parameter combination corresponding to the smallest objective function value is our LSE. We rerun the routine for 30 realizations of the incidence rates and plot the correlation between $\mu$ and $\nu$ (figure 4.2).

There is a clear reciprocal relationship between $\mu$ and $\nu$ for the first scenario. However, when the incidence rate reaches its asymptotic height or the data is more complete, the fluctuations in the data do not affect the estimates much since the chance that the height in the tail would be shifted by random noise is small. The estimate of the birth rate $\alpha$ is always stable and precise.

Accuracy of estimates when incorporating covariates into the biological parameters will be further discussed in the data analysis sections. We begin with the data analysis for never smokers.
4.4 Estimation of baseline biological parameters based on never smokers data

There are several factors, namely, mutagen sensitivity, DNA repair capacity, smoking history, and gender, involved in this carcinogenesis model. It is not easy to estimate all the parameters simultaneously with good precision for a non-homogeneous population using combined incidence data alone. Moreover, we are also interested in the effect of genetic susceptibility on never smokers. Following our original plan, our first step
will be to estimate the full model with no smoking assumed in order to obtain the baseline parameter estimates for current smokers data. The CPS-II (1982-1988) data provides age-specific lung cancer incidence rates for never smokers for females and males separately. There were no genetic susceptibility measurements available then. However, the controls (healthy people) in the M.D. Anderson data can serve as a source to approximate the distribution of genetic susceptibility among never smokers. In the white population, we assume two levels (1/2, 1) of DNA repair capacity (DRC) using the median value of DRC among controls as a cutoff. The whole population is divided into two subpopulations: good-DRC with DRC level 1 and poor-DRC with DRC level 1/2. A finer partition of the whole population was not considered due to the limitations of the sample size. We will run LSE using the incidence data collected for the whole population to estimate the baseline biological parameters $\mu_0, \nu_0, \alpha_0,$ and $\beta_0$.

### 4.4.1 Estimation based on a mixed dataset

In order to estimate the impact of DRC, the ideal data would be the stratified incidence rates data for good/poor DRC subpopulations. However, for the CPS-II data, it is unrealistic to ask for thousands of people's DRC to be measured. Whether mixed incidence data is sufficient to infer the DRC coefficient should be examined first. Unlike the response function set-up for never smokers at the beginning of this chapter, we consider more general circumstances, modify the response function, and allow the
two coefficients $a_1$ (the DRC effect in the first transition rate) and $a_2$ (the DRC effect in the second transition rate) to be different. The first concern is whether the signal it gives is real. To verify the possibility of giving a false signal, we create a scenario with no DRC effect. We compute five-year bin incidence rates without adding any noise and estimate parameters by LSE. The simulation results confirmed our guess that the mixed incidence data might produce a false signal.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$\mu$</th>
<th>$\nu$</th>
<th>$\alpha$</th>
<th>$a_1$</th>
<th>$a_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimized</td>
<td>9.98e-8</td>
<td>0.26</td>
<td>0.36</td>
<td>1.75</td>
<td>0</td>
</tr>
<tr>
<td>Assumed $a_1 = a_2$</td>
<td>9.7e-8</td>
<td>0.5821</td>
<td>0.36</td>
<td>0.0413</td>
<td>0.0413</td>
</tr>
<tr>
<td>True parameter value</td>
<td>1e-7</td>
<td>0.6</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.2: Least Squares Estimation for Mixed Incidence Data Without Noise

As noticed, there are slight deviations in the estimates from the true values even in the absence of noise. We would expect a bigger error in the estimates than those in the table above when dealing with real data with noise.

This simulation study showed us a potential problem associated with estimation based on combined incidence data alone. The false signal was caused by the absence of information on relative risks between the two subpopulations. The relative risk between two subpopulations can be found in the case-control data. We propose the use of the ratio of the cumulative risk of developing lung cancer between the ages of 30 and 75 (forty-five year age range). Age-specific (five year bin) ratios were our first
target since they contain richer information than ratios of cumulative risks and thus can reveal more aspects of carcinogenesis. The limited sample size forced us to give up partitioning never smoker patients in the M.D. Anderson data into smaller groups with five year age ranges.

Therefore, instead of minimizing the sum of squares of the difference between the predicted incidence and the observed incidence alone, we suggest adding one more component, which contains the information on the cumulative risks for two subpopulations, to the objective function. Setting one group as the baseline, we compute the square of the difference between the cumulative risk of the second group predicted by the model and the cumulative risk of the second group computed based on the ratio.

\[ f_2 = \frac{((P(35 \leq T \leq 80|\text{DRC level 2}) - P(35 \leq T \leq 80|\text{DRC level 1}) \times R_{21}) \times 10^5)^2}{(P(35 \leq T \leq 80|\text{DRC level 2}) \times 10^5)}, \]

where

\[ R_{21} = \frac{P_2}{P_1} = \frac{P(35 \leq T \leq 80|\text{DRC level 2})}{P(35 \leq T \leq 80|\text{DRC level 1})}. \]

The population size 10^5 is taken arbitrarily here and can be replaced with other appropriate number when the numbers of subjects in the datasets or the cumulative risks are changed. The new objective function is \( f_1 + f_2 \), where \( f_1 \) is the Chi-square sum defined earlier. We can put different weights on the ratio information part. However we do not have time to pursue this direction. Also, in order to improve the smoothness of the objective function and reduce some unexpected large objective
function values, a logarithmic transformation was applied. The issue of what type of transformation is best for the optimization will not be investigated here either. Again, we estimate the parameters under the same simulation scenario by using the new objective function and no perturbation for overall incidence rates. The new estimates clearly better exclude false signals.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$\mu$</th>
<th>$\nu$</th>
<th>$\alpha$</th>
<th>$a_1$</th>
<th>$a_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimized $a \in [0, 2]$</td>
<td>1e-7</td>
<td>0.6</td>
<td>0.36</td>
<td>0</td>
<td>0.000</td>
</tr>
<tr>
<td>$a \in [-1, 2]$</td>
<td>1.0015e-7</td>
<td>0.5991</td>
<td>0.36</td>
<td>0.0059</td>
<td>-0.0059</td>
</tr>
<tr>
<td>$a \in [-1, 2] \ a_1 = a_2$</td>
<td>1e-7</td>
<td>0.6</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True parameter value</td>
<td>1e-7</td>
<td>0.6</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4.3: LSE with DRC ratio information and population size $10^7$

Apparently, the new objective function does a better job in finding the true parameter values. There is still some negative correlation between $\mu$ and $\nu$ in the case when $a_1$ and $a_2$ are allowed to take negative values. But imposing the equality $a_1 = a_2$ improves the precision of the final estimates. Next, we need to consider situations when the observed number of lung cancer cases is subject to Poisson random noise. One of our simulation studies shows (figure: 4.3) that there are large fluctuations in the estimates of $\mu$ and $\nu$. The trade-off between $\mu$ and $\nu$ remains. But the coefficients of $a_1$ and $a_2$ remain stable: both are nearly zero.

Similar to the simulation comparison in the beginning of this chapter, we will show
Figure 4.3: Estimates of parameters when missing incidence data and noise are both present, $a_1 = a_2$ assumption is enforced. The estimates of $a_1$ and $a_2$ are nearly zero. The two lower plots are based on 10 runs.
how more complete observed incidence data will provide extra information leading
to better precision of the estimates. According to the simulation results, even in the
presence of Poisson random noise in the observed cases, the estimates are stable and
precise (figure: 4.4).

Furthermore, we will show that this assumption $a_1 = a_2$ is necessary and practical
even in the presence of a real signal, i.e. $a_1, a_2 > 0$. This time, we assume nonzero
DRC coefficients $a_1 = 1$ and $a_2 = 0.3$ and use the new modified objective function $f_1 + f_2$. The effect of DRC on the first and second transition rate are always confounded
with the other (see figures 4.5 and 4.6).

Frequently both estimates fall into a region between the two true parameter val-
ues. For survival probability prediction and signal detection purposes, it might be
reasonable to assume identical coefficients for the two transition rates. For this reason
we assume equal coefficients for the same risk factors in the first and second mutation.

Combining the two levels of the DRC subpopulations among different age groups
will be difficult when no such information is available. One possibility is to compute
the probability of survival to a certain age for each subpopulation and then calculate
the ratio of the number of people alive to that age among each subpopulation. The-
oretically, this will yield the best guess of the composition of the two subpopulations
in different age groups. But some uncertainty arises when competing death causes
are taken into consideration. The influence of competing causes becomes stronger as
people get older. To the best of our knowledge, there is no evidence in the literature
Figure 4.4: Estimates of parameters with more complete data and noise, $a_1 = a_2$ assumption is enforced. The estimates of $a_1$ and $a_2$ are nearly zero. The two lower plots are based on 10 runs.
Figure 4.5: Density of estimates of $a_1$ and $a_2$, true $a_1 = 1$, $a_2 = 0.3$

correlation between $a_1$ and $a_2$

Figure 4.6: Dependence between estimates of $a_1$ and $a_2$

that competing risk has been considered in the carcinogenesis setting. Therefore, this problem may not be resolved completely and correctly at this stage. Since we use
the median of the controls as the cutoff and the lung cancer incidence among never smokers is extremely small, it is reasonable to assume that the number of people with poor-DRC is equal to the number of people with good-DRC in the white population. Hence we simply mix the two subpopulations half by half in each age group in our data analysis and simulation study. The implications and problems of the assumption of a one-to-one ratio will be discussed in the last chapter.

4.4.2 Estimation of the effect of DRC on never smokers lung cancer development

The choice for the true death rate $\beta_0$ averaged over a lifetime is difficult. We take some information from Heidenreich's paper ([12]) as guideline for the selection of $\beta_0$. We matched the same magnitude of $\nu$ in females and males as that in Heidenreich's paper ($\nu \approx 0.01$). The lung cancer incidence data is taken from Appendix 4 and 5 in the CPS-II report. Using the estimation approach described above and incorporating the ratio of cumulative risk of people with poor DRC and good DRC from the M.D. Anderson case-control data, we run the analysis for white female and male never smokers and obtained the following results.

The 95% C.I.s are constructed based on 100 runs. Each time we resample the numbers of lung cancer cases in each age group according to a Poisson distribution of counts, we compute the corresponding incidence rates. Then we treat the incidence
<table>
<thead>
<tr>
<th>Setting</th>
<th>$\nu$</th>
<th>$\mu$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$a_1$</th>
<th>DRC ratio</th>
<th>95% C.I. for $a_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.0119</td>
<td>7.8557e-7</td>
<td>1.4925</td>
<td>1.4</td>
<td>3.3569</td>
<td>6.2</td>
<td>(2.99, 4)</td>
</tr>
<tr>
<td>Male</td>
<td>0.0114</td>
<td>2.7638e-7</td>
<td>1.115</td>
<td>1</td>
<td>3.0211</td>
<td>5.5</td>
<td>(2.7, 3.9)</td>
</tr>
</tbody>
</table>

Table 4.4: LSE and C.I. for never smokers

rates as observed and rerun the estimation for all parameters and coefficients. The estimates of $a_1$ in both genders are positive. The genetic factor has a very similar effect on female and male never smokers. The slightly higher lung cancer incidence among men can mainly be explained by a slightly higher cell growth rate. The difference in the effect of DRC will not cause an obvious difference between the two genders.

### 4.5 Current smoker parameter estimation

In the never smokers case, there is only one risk factor in the response functions. When dealing with current smokers, we face a more heterogeneous population. The CPS-II data provide age-specific mortality for subpopulations stratified by smoking intensity (20 or 40 cigarettes per day) and smoking duration (30 to 34 years, 35 to 39 years, 40 to 44 years, 45 to 49 years and more than 50 years). This is the ideal dataset. The other mortality tables from the CPS-II study only provide the overall death rates for current smokers. These are age-specific death rates for current smokers with variations in smoking initiation age and smoking intensity. To compute the overall incidence rate, we need the distributions of these latter two variables.
Three levels of smoking intensities: 10, 20 and 40 cig/day and three initiation ages: 20, 25, 30 for females and 15, 20 and 30 for males are assumed. The distributions are as follows:

<table>
<thead>
<tr>
<th>Initiation age</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>0.3</td>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td>Initiation age</td>
<td>15</td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Males</td>
<td>0.35</td>
<td>0.5</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Table 4.5: Distribution of smoking initiation age for current smokers

<table>
<thead>
<tr>
<th>Smoking Intensity cigarette per day</th>
<th>10</th>
<th>20</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.35</td>
<td>0.5</td>
<td>0.15</td>
</tr>
<tr>
<td>Male</td>
<td>0.15</td>
<td>0.55</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table 4.6: Distribution of smoking intensity for current smokers

There are numerous ways to construct response functions of risk factors and biological parameters in the presence of cigarette smoking since the exact mechanisms behind the transition and promotion processes and their association with cigarette smoking and genetic predispositions are unknown. Although we have the estimate of coefficient $a_1$ for DRC for never smokers, I do not use this estimate in the response function when cigarette smoking is present. One can argue that exposure to higher doses of carcinogen might change the way DRC affects $\nu$ and $\mu$. We choose two simple
forms of the response functions:

**Response function 1**

In the presence of cigarette smoking,

\[ \nu = \nu_0(1 + a_2)(1 + a_4(1 - DRC)) \]

\[ \mu = \mu_0(1 + a_3)(1 + a_4(1 - DRC)) \]

\[ \alpha = \alpha_0(1 + a_5 \times (\log(sd) - a_6)) \]

\[ \beta = \beta_0(1 + a_5 \times (\log(sd) - a_6)). \]

In the absence of cigarette smoking, where \( a_1 \) was estimated from never smokers data

\[ \nu = \nu_0(1 + a_1(1 - DRC)) \]

\[ \mu = \mu_0(1 + a_1(1 - DRC)) \]

\[ \alpha = \alpha_0 \]

\[ \beta = \beta_0. \]

**Response function 2**

In the presence of cigarette smoking,

\[ \nu = \nu_0(1 + a_2)(1 + a_4(1 - DRC)) \]

\[ \mu = \mu_0(1 + a_3)(1 + a_4(1 - DRC)) \]

\[ \alpha = \alpha_0(1 + a_5(\log(sd) - a_6)) \]

\[ \beta = \beta_0(1 + a_5(\log(sd) - a_6)). \]
In the absence of cigarette smoking, where $a_1$ was estimated from never smokers data

$$\nu = \nu_0(1 + a_2)(1 + a_1(1 - DRC))$$

$$\mu = \mu_0(1 + a_3)(1 + a_1(1 - DRC))$$

$$\alpha = \alpha_0$$

$$\beta = \beta_0.$$  

In response function 1, the two terms $(1 + a_2)$ and $(1 + a_3)$ represent the influence of cigarette smoking on the first transition and second transition rates in the TSCE model. However, in response function 2, we assume that the mechanism of lung cancer for current smokers is different from the mechanism of lung cancer for never smokers. Hence $\nu(1 + a_2)$ and $\mu(1 + a_3)$ are the baseline transition rates for lung cancer in current smokers. In this case, we do not consider a smoking effect in transition rates because of parameter non-identifiability.

We will continue to use the LSE. Let us restate the objective function:

$$f = \sum_{i}^{m} \frac{(O_i - E_i)^2}{E_i} + \frac{((P_{11} - P_{12} \times R_{12}) \times 10^3)^2}{P_{11} \times 10^3} + \frac{((P_{21} - P_{22} \times R_{22}) \times 10^3)^2}{P_{21} \times 10^3}.$$

In the first term, $O_i$ is the number of observed deaths due to lung cancer according to the CPS-II data, $E_i$ is the expected numbers of death predicted by the TSCE model under a given parameterization, and $m$ is the number of cells in all the tables. The other two terms correspond to risk ratio information for people who smoke 20 or 40 cig/day. In the second term, $P_{11}$ and $P_{12}$ are the probabilities that a person
with poor/good DRC develops lung cancer between the ages of 35 to 80 predicted by the TSCE model given a smoking intensity of 20 cigarettes per day, and $R_{12}$ is the observed ratio of the two probabilities approximated from the case-control data. Using a similar notation, $P_{21}$, $P_{22}$, and $R_{22}$ are the probabilities and their ratio when smoking intensity is 40 cigarettes per day. The population size in the last two terms drops from $10^5$ in the never smokers case to $10^3$ since the probability of developing lung cancer in a lifetime for current smokers is as 50 to 100 times as that for never smokers. Reducing the population size allows the death counts in the last two terms, such as $P_{11} \times 10^3$, to be comparable with those in the first term (Chi-square sum).

How did we obtain $R_{12}$ and $R_{22}$, the ratios of the cumulative probabilities for the two subpopulations from the case-control data? Similar to the never smokers case, we only consider two levels of DRC measurements among the white population: good DRC (level 1) and poor DRC (level 1/2). People with good (efficient) DNA repair capacity can fix more DNA errors than people with poor DRC. We need to find the median cutoff and divide the whole population into two subpopulations. For current smokers, two levels of smoking intensities are considered: 20 cigarettes per day and 40 cigarettes per day, since the CPS-II data only provides stratified data for these two levels. In order to obtain stable estimates of the ratios, we pool all the current smokers and recent quitters with smoking intensity ranging from 15 to 25 cigarettes per day together and treat them as current smokers with smoking intensity
20 cigarettes per day. Similarly, we group all the current smokers and recent quitters with smoking intensity 30 to 40 cigarettes per day and treat them as current smokers with smoking intensity 40 cigarettes per day. The median of the DRC levels among the white population is approximated by the median of DRC levels among controls with appropriate smoking intensity. We have a median of 9.12 for females and 8.68 for males who smoke 20 cigarettes per day and a median of 8.33 for females and 8.24 for males who smoke 40 cigarettes per day.

Using the median cutoff, we divide the patients with corresponding smoking intensity into two groups: patients with good DRC and patients with poor DRC. The ratios $R_{12}$ and $R_{22}$ of interest are estimated by the ratio of the number of patients with good DRC versus the number of patients with poor DRC. The ratios $R_{12}$ and $R_{22}$ for both genders and pooled data are summarized below:

<table>
<thead>
<tr>
<th>Smoking intensity</th>
<th>20 cig/day</th>
<th>40 cig/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>Females</td>
<td>5.83</td>
<td>2</td>
</tr>
<tr>
<td>Males</td>
<td>2</td>
<td>1.25</td>
</tr>
<tr>
<td>Total</td>
<td>2.94</td>
<td>1.89</td>
</tr>
</tbody>
</table>

Table 4.7: Ratio between poor DRC and good DRC among lung cancer patients

Apparently, the ratio for female smokers of 20 cigarettes per day is extremely high, and even higher than the ratio estimated for never smokers. If we examine these
ratios across the two genders, we see a big difference for smokers with 20 cig/day. It seems doubtful that this gender difference is real since the ratio for smokers with 40 cigarettes per day is relatively close for both women and men. Moreover, for female former smokers of 20 cig/day, it appears that the figure is closer to the figure of current male smokers. At this point in the analysis, we ignore the gender differences and use ratios built with the pooled (females and males) data.

The estimates are based on all the counts in Appendix 5, 6 in Chapter 4 [31] except cells with a single death count. It seems that the exclusion of the single-death-count cells stabilizes the optimization procedures. We refer to the data after the truncation of the single death counts as “cleaned data.” Weight adjustments were made for some terms in the objective function $f_1$. The weight of death counts for people smoking more than 50 years decreases from 1 to 0.8. Using combined odds ratio information, we obtained the following estimates for female and male current smokers. We also calculated the overall age-specific incidence rates for current smokers given the initiation age and smoking intensity distributions in the previous tables. We refer to this probability of developing lung cancer by age of 75 as $P_{75}$.

The plots of fit of the response function 1 are from figure 4.7 to figure 4.12.

The estimate of $a_3$ is negative in response function 1 for both females and males. There is a hypothesis that the predominant lung cancer type for smokers is different from that for non-smokers. If this is true, the mechanism might be different as well, implying that the interpretations of $a_2$ and $a_3$ should change. The baseline transition
<table>
<thead>
<tr>
<th>Parameters</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$a_6$</th>
<th>$P_{75}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.7385</td>
<td>-0.7079</td>
<td>20</td>
<td>0.5009</td>
<td>1.60</td>
<td>0.03</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>(-0.90, 2.24)</td>
<td>(-0.99, 0.54)</td>
<td>(5.31, 50)</td>
<td>(0.35, 0.66)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>-0.0039</td>
<td>-0.99</td>
<td>20</td>
<td>0.1671</td>
<td>-1.24</td>
<td>0.026</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>(-0.69, 1.37)</td>
<td>(-0.9999, -0.78)</td>
<td>(7.09, 50)</td>
<td>(0.08, 0.23)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.8: LSE for current smokers using cleaned data and response function 1

Figure 4.7: Fit of the TSCE model for current female smokers 20 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations.
Figure 4.8: Fit of the TSCE model for current female smokers 40 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations.

rates for current smokers lung cancer are modified to $\nu(1 + a_2)$ and $\mu(1 + a_3)$. These baseline rates should also be applied to the period before smoking starts.

This set-up is investigated in the response function 2. Repeating the same procedure, we obtain estimates for response function 2 and predicted probability of developing lung cancer by age 75, $P_{75}$.

The overall risk of developing lung cancer by age of 75 is around 16% for males
Figure 4.9: Fit of the TSCE model for current female smokers 20 cig/day and 40 cig/day using response function 1.

and 9.5% for females according to the figures provided by Peto and Doll [25]. It seems that response function 2 gives a better prediction. Also, response function 2 yields smaller objective function values than response function 1 for both males and females. Therefore, we place greater trust in the mechanism proposed in response function 2 and make inference and predictions on this setting in the next chapter.
Figure 4.10: Fit of the TSCE model for current male smokers 20 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations.

<table>
<thead>
<tr>
<th>Setting</th>
<th>( a_2 )</th>
<th>( a_3 )</th>
<th>( a_4 )</th>
<th>( a_5 )</th>
<th>( a_6 )</th>
<th>( P_{75} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>3.5279</td>
<td>-0.3410</td>
<td>11.4619</td>
<td>0.3771</td>
<td>2.28</td>
<td>0.06</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>(0.15, 7.42)</td>
<td>(-0.74, 1.68)</td>
<td>(7.97, 15.14)</td>
<td>(0.21, 0.65)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>3.6670</td>
<td>-0.6899</td>
<td>15.3117</td>
<td>0.1655</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>95% C.I.</td>
<td>(0.23, 13.73)</td>
<td>(-0.91, -0.19)</td>
<td>(4.23, 16.76)</td>
<td>(0.07, 0.31)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.9: LSE for current smokers using cleaned data and response function 2
Figure 4.11: Fit of the TSCE model for current male smokers 40 cig/day using response function 1. The overall incidence is the incidence for smokers with different smoking intensities and durations.
Figure 4.12: Fit of the TSCE model for current male smokers 20 cig/day and 40 cig/day using response function 1.
Figure 4.13: Fit of the TSCE model for current female smokers 20 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations.
Figure 4.14: Fit of the TSCE model for current female smokers 40 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations.
Figure 4.15: Fit of the TSCE model for current female smokers 20 cig/day and 40 cig/day using response function 2.
Figure 4.16: Fit of the TSCE model for current male smokers 20 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations.
Figure 4.17: Fit of the TSCE model for current male smokers 40 cig/day using response function 2. The overall incidence is the incidence for smokers with different smoking intensities and durations.
Figure 4.18: Fit of the TSCE model for current male smokers 20 cig/day and 40 cig/day using response function 2.
Chapter 5

Conclusions and discussions

5.1 Conclusions

We applied the TSCE model to CPS-II age-specific lung cancer death rates data combined with DRC information collected from a case-control study. The influence of cigarette smoking and genetic factors in lung cancer development was assessed through parameter estimations. We seem to be the first to do this. We tried two simple response functions to describe the relation between risk factors and biological parameters. The main conclusions we have reached from this study are:

- Cigarette smoking is the major risk factor for lung cancer development since cigarette smoking increases the initiated cells' net proliferation rate and shortens the waiting time for the rise of tumors.

- The effect of cigarette smoking on transition rates is minor compared to its
effect on proliferation rates.

- Genetic susceptibility, measured by DRC in this case, is the secondary risk factor for lung cancer development.

- There may be a gender difference in terms of how cigarette smoking and genetic factors alter the initiation and promotion of cancerous cells.

- There may be a mechanism difference between the predominant types of lung cancer in smokers and never smokers.

Using the estimates from response function 2, we can compute the risk of developing lung cancer for never smokers and current smokers stratified by DRC levels. Even for the best cases, i.e. people with good genetic heritage, cigarette smoking will increase their chance of having lung cancer by about 10 fold. The predicted cumulative risks of developing lung cancer by age 75 are outlined in table 5.1:

<table>
<thead>
<tr>
<th>Response function 2</th>
<th>never smokers</th>
<th>current smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>good DRC</td>
<td>poor DRC</td>
</tr>
<tr>
<td>Females</td>
<td>0.001</td>
<td>0.0068</td>
</tr>
<tr>
<td>Males</td>
<td>0.0012</td>
<td>0.0072</td>
</tr>
</tbody>
</table>

Table 5.1: Cumulative risk of developing lung cancer by age 75 among different subpopulation. For current smokers, it is based on initiation age 15 and smoking intensity 20 cig/day.
For male never smokers, we conclude that genetic susceptibility can alter the chances of lung cancer onset. That is, DRC is an important risk factor for non-smokers. The cumulative risk for people more susceptible to lung cancer is six times higher than that for people with good DRC. Insufficient DNA repair increases the frequency of deleterious mutations, speeds up the two transition processes, and thus results in a higher incidence rate among people with poor DRC. Regardless of DRC level, male current smokers uniformly have a much higher risk of developing lung cancer than never smokers. Cigarette smoking is certainly the main cause of lung cancer. The baseline level of the first transition rate for smokers' lung cancer is larger than that for never smokers’ lung cancer. This might be due to the effect of cigarette smoking. The baseline of the second transition rate for smokers' lung cancer is surprisingly smaller than that for never smokers’ lung cancer. This implies that the predominant lung cancer type among current smokers is rarely seen among never smokers. The high lung cancer incidence among smokers is mainly caused by the increased net proliferation rate, which is the major force of lung cancer carcinogenesis. The role DRC plays in the presence of smoking is more dramatic than its effect for never smokers. The coefficient jumps from 3 for never smokers to 15 for current smokers. DRC becomes less important among heavy smokers due to the overwhelming effect of cigarette smoking. This pattern was also observed in the case-control data. The ratio of good DRC versus poor DRC among patients decreases from 2 to 1.25 when the smoking intensity increases from 20 cig/day to 40 cig/day.
Similar to male never smokers, poor DRC can enhance transition rates for female never smokers. The cumulative risk for never smokers with poor DRC is around seven times that for never smokers with good DRC. DRC is a more crucial factor for female never smokers than for male never smokers. For female current smokers, the DRC effect is strong for light smokers and diminishes for heavy smokers. Cigarette smoking largely increases the proliferation rate. Similar to male smokers, the baseline level of the first transition rate is higher for smokers than for never smokers. However, the baseline of the second transition rate for smokers is not significantly lower than that for never smokers. The true baseline of the transition rates may be confounded with the effect of cigarette smoking. With current data and the TSCE model, the confounding problem cannot be solved. The coefficient corresponding to the cigarette smoking effect in the proliferation rate for females is larger than that for males. It suggests higher smoking intensity might push the risk of developing lung cancer even higher for women than for men because the net proliferation rate increase is larger for women.

In the case-control data, it seems that poor levels of DRC leave women more susceptible to lung cancer than men. This mechanism is not well understood. Across all lung cancer incidence or mortality rate data from the CPS-II study, the Texas lung cancer registry and the SEER data, one consistently present feature is that women have a slightly higher cumulative risk of having lung cancer before the age of 30. Among young patients, cigarette smoking is hardly to be blamed for the onset of
lung cancer since the duration of smoking is still short and its effect can be negligible. Therefore genetic susceptibility may be the main drive, confirming the conclusion from our study that DRC has a slightly bigger influence on lung cancer development in females or when the carcinogenesis exposure is low. Additional independent evidence pointing to this conclusion comes from the distribution of different types of lung cancer among men and women, never smokers and ever smokers. According to the CNN health/library website on lung cancer facts [19], there are two types of lung cancer: small cell lung cancer and non-small cell lung cancer. Within non-small cell lung cancer, there are three kinds of lung cancer, classified according to the primary tumor site: adenocarcinoma, squamous cell carcinoma and large cell lung cancer. Interestingly, “Small cell lung cancer almost surely appears only among smokers. Adenocarcinoma is common among women and never smokers and squamous cell carcinoma is more common among men.” One hypothesis would be that women are more vulnerable to lower doses of carcinogen exposure because the mechanism of cancerous cell development for most women is similar to that of never smokers. To the best of our knowledge, no paper makes use of this association in an attempt to propose different pathways of carcinogenesis for never smokers and current smokers in lung cancer.

All our data analyses are based on ratio information combining female and male data. We reran the entire procedure for ratios computed for each gender separately and found the coefficients slightly changed. The major conclusions remain the same.
When combining incidences of people with different DRC levels, we mixed the incidence rates of two subpopulations half by half. For advanced age groups, we also tried slightly increasing the proportion of people with good-DRC. The coefficients corresponding to DRC and cigarette smoking increase but the conclusions do not change either. In our model, the effect of cigarette smoking on transition rates is not as large as its effect in proliferation rates. This is in agreement with findings of former studies by Moolgavkar’s group ([8] and [9]).

5.2 Problems and future research direction

The first unsolved problem is: Why is there a gender difference and a mechanism difference? The second problem revealed in the CPS-II data and British physician study [30] is that lung cancer incidence is not common among old people (age greater than 80) and lung cancer incidence is much smaller among the heaviest smokers (more than 40 cigarettes per day). The lower incidence in the highest age group might be explained by the overturn hypotheses suggested by Pompei and Wilson ([27] and [26]). These hypotheses are supported not only by the uniform overturn in incidence curves observed in all human cancer data around the world but also by the overturn of incidence curves observed in mouse tumor data. The slower growth of breast cancer among old patients, and the greater aggressive invasion and metastasis among young patients back Pompei’s conjecture ([24] and [33]). If this is true, a series of research projects could be developed to study the overturn age and factors which trigger the
overturn.

The causal relationship between DRC and lung cancer can only be confirmed through a prospective study, a study designed to follow up a group of controls with different levels of DRC over several years. If the positive association between the DRC level and the presence of lung cancer is significant, then we can conclude with more confidence that DRC is a cause of lung cancer.

Of course the most important question is still how cancerous cells evolve. In this thesis, our main goal is to study lung cancer incidence on a population basis. From an epidemiological point of view, identifying major risk factors and predicting future incidence are the major concerns. The non-identifiability problem would not undermine our conclusions if assuming one or two parameter values beforehand would be acceptable. However, for oncologists and physicians studying the mechanism of carcinogenesis, looking for prevention strategies to slow down cancer onset, and finding treatments to control or contain the invasion of tumors seem more relevant. There is a long way to go from what current stochastic carcinogenesis models can offer to a complete understanding of the mechanisms of carcinogenesis for various types of cancer. However, using incidence data alone is definitely not sufficient or appropriate to assess parameters associated with intermediate or latent stage cancers [30]. The intermediate stage and preclinical stages hold the key to cancer prevention and maybe even the key to cancer treatment. Prolonging the latent time might lead to a cancer free life. Both prevention and treatment depend on research regarding the dynamic
changes of initiated and cancerous cells and their interaction with the environment.

To demonstrate the dynamic changes of initiated cells, animal models are indispensable. New discoveries in animal experiments have forced researchers in the area of stochastic modeling to revise their models and derive better mathematical models or to propose more reasonable hypotheses of carcinogenesis. One research group in the United Kingdom compared two multistage models, namely the TSCE model and multi-stage and multi-path model, with a new stochastic multigate model. They propose their new model after studying induced tumors in p53 deficient and wild type mice. Mao et al. [18] ruled out the simple multi-stage model and suggested that multi-stage multi-path and multi-stage multigate models were more plausible based on their data. In their multi-gate model, the p53 gene is called the gatekeeper. A mouse with two p53- genes has larger transition rates for several transition stages and in order to pass the gates (turn from p53++ to p53- - ) at least two mutation events are required. They explained that “this is a genetic instability model and it requires that p53 inactivation be a destabilizing event.” Therefore, this animal model supports the genome stability theory, one of the most popular carcinogenesis theories. Carcinogenesis models based on genome integrity are worth future exploration. Mao’s computation for the number of tumors at a given time was simulation based due to the complexity of his model.

During my study of the TSCE model, I also attempted to refine an old stochastic carcinogenesis model and to develop my own model to resolve the difficulties occurring
when using a multi-stage single path model to explain real data. I will describe some of my preliminary findings.

There are six traits commonly recognized as the hallmarks of cancer. It appears that the number of six agrees with Armitage’s model. However, the order of these traits can vary from case to case, and each stage might include more than one events such as mutation or cellular change. Hence, to map these six traits to Armitage’s six-stage model is not straightforward. In fact, some researchers suggested two to ten distinct stages for cancer development depending on the cancer type [18]. Despite the mystery of the mechanism underlying cancer, most biologists accept the evolution theory of cancerous cells.

Inspired by the evolution theory of cancerous cells, I define the fitness of the abnormal cells as the malignancy of the cells and build a new model called the beta-shift model. The purpose of this model is to capture dynamic changes in the malignancy of abnormal cells under simple assumptions.

Unlike the N-A and the TSCE models, we do not use a fixed number of stages (two or six). Instead we assign a malignancy level $m$ to a cell on a continuous scale with range between 0 and 1. The malignancy value for normal cells is zero, and for completely malignant cells it is one. At time zero, the assumption is that every cell is normal and there are about $10^6$ to $10^7$ candidate cells (usually stem cells) for transition. When a cell jumps, it experiences a malignancy change. The waiting time for a single cell to gain a more selective advantage follows the exponential distribution.
The increment of malignancy, $m_2 - m_1$, is $\text{Beta}(a, b(1 - m_1))$ distributed. After the jump, this first abnormal cell can divide slowly and wait for opportunities to jump again. Cells with lower fitness cannot compete with the winner cell(s) and stop dividing and jumping. Only the activities of the winner cells are of interest and recorded in this model. The offspring of the winner cells share the same malignancy, so they compete for the next change. This process repeats in the same fashion until one winner cell's malignancy level passes a certain threshold. (This is illustrated in figure 5.1.)

![Graph showing evolutionary path]

Figure 5.1: Evolutionary path for the first rise of a cancerous cell (different paths for different individuals)

The beta-shift model enjoys the following plausible properties:

(1) It allows for heterogeneity among intermediate (abnormal) cells.

(2) The paths of carcinogenesis for different patients can vary.
(3) There is no fixed number of transition stages.

(4) The focus is on the winner cells’ expansions and jumps.

(5) It allows for flexibility in adding new discoveries to the model.

(6) The “diagnosed” tumor can have variable malignancy.

As the first winner cell waits for its next jump, it proliferates according to a birth process with rate \( \lambda \) and forms a clone. The waiting time for a clone to jump to next level has been well studied ([14]). The survival function is given by

\[
S(t) = \frac{1}{(1 - \alpha) + \alpha e^{\lambda t}},
\]

where \( \alpha \) is the probability that one daughter cell mutates during cell division and \( \lambda \) is the division rate.

The distribution of the time for the first cell to pass a given threshold is mathematically intractable. We rely on simulation to approximate this distribution. The hazard function under the beta-shift model increases exponentially with time \( t \). However, in the log scale (figure 5.2), the incidence rate shows a bi-phase feature. We see a flat line (no cancers observed in the early stages), then a sharp increase, but with a slope that slightly decreases as time increases. This feature does not agree with the Texas Cancer Registration Data. With this data, the slope of the incidence rate gradually increases in time.

Although the model under the six assumptions was not a good fit to the incidence data, changing some assumptions might help. I examined hazard functions for all
Figure 5.2: Simulation based incidence rates built on the beta-shifted model. Red dots are observed Texas lung cancer incidence rates and blue dots are simulated incidence rates based on 10000 samples.

common continuous distributions. It appears that only the double exponential and normal distributions have curve shapes similar to that of the observed cancer incidence curve. My simulation incidence curve is closer to the hazard function of the gamma distribution. This is not surprising since the sum of several waiting times (exponential r.v.) is gamma distributed. Because of the Central Limit Theorem, a large number of small steps (jumps back and forth many times) lead to an approximately normally distributed waiting time.

Finally, for testing these carcinogenesis models, incidence data as highly summa-
Figure 5.3: Hazard function for the normal distribution with normal censoring

Rized numbers cannot identify all the feature parameters we mentioned above. More specific, detailed tumor data such as the tumor size and the proportion of different malignancy levels would be necessary to estimate these parameters.

In addition to changes in assumptions, the need to search for a faster sampling algorithm and a more efficient programming language is also urgent. A better set of tools could reduce simulation times to a realistic scale without sacrificing the huge sample sizes required for constructing probability functions such as the survival function, the CDF and the hazard function. We believe the beta-shift model will be useful in resolving some of our questions if suitable assumptions on the biological process and data sets are applied.
Figure 5.4: Hazard function for the double exponential distribution
Bibliography


Appendix A

Assumptions and PGF derivation of hazard function

The following is copied from Tan’s book ([29]). We denote the number of normal stem cells, intermediate cells and malignant cells by \( N(t) \), \( I(t) \) and \( Z(t) \), respectively. For notational convenience and clarity, we first need introduce some definitions, notation and assumptions that will be used in our discussion throughout the remainder of this thesis.

**Assumptions:**

1. At the starting point \( t_0 \) (usually \( t_0 = 0 \)), there are \( N_0 \) NCs.

2. During \( [t, t + \Delta t] \), the probability that a normal stem cell at time \( t \) yields one normal stem cell and one initiated cell is \( \nu(t) \Delta t + o(\Delta t) \), where \( \lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0 \).

   Similarly, the probability that an initiated cell yields one IC and one MC during
\[ [t, t + \Delta t] \text{ is } \mu(t) \Delta t + o(\Delta t). \]

3. The ICs are assumed to follow a nonhomogeneous birth-death process with birth rate \( \alpha(t) \) and death rate \( \beta(t) \).

4. The time for an MC to develop into a clinically detectable cancer tumor is negligible.

5. The birth-death processes and the mutation processes are independent of each other and each cell proceeds through the above processes independently of other cells.

In order to find the probability distribution of time required for a normal stem cell to develop into a tumor, we need to proceed to find the age specific incidence function through probability generating function (PGF).

**Definitions and Notation**

0. Let \( T \) be the time required for a single malignant cell to arise starting with a pool of \( N_0 \) normal stem cells.

1. Let \( \psi(t_0, t) \) be the joint PGF of \([I(t), Z(t)]\) given \([N(t_0) = N_0, I(t_0) = Z(t_0) = 0]\).

\[
\psi(t_0, t) = \psi(t_0, t; x_2, x_3) = \sum_{i_2} \sum_{i_3} x_2^{i_2} x_3^{i_3} P_1(i_2, i_3; t_0, t);
\]

where \( P_1(i_2, i_3; t_0, t) = P(I(t) = i_2, Z(t) = i_3 | N(t_0) = 1, I(t_0) = Z(t_0) = 0). \)
2. Let $\phi(s,t)$ be the joint PGF of $[I(t), Z(t)]$ given $[I(s) = 1, N(s) = Z(s) = 0]$.

$$
\phi(s,t) = \phi(s,t; x_2, x_3) = \sum_{j_2} \sum_{j_3} x_2^{j_2} x_3^{j_3} P_2(j_2, j_3; s, t);
$$

where $P_2(j_2, j_3; s, t) = P(I(t) = j_2, Z(t) = j_3 | I(s) = 1, N(s) = Z(s) = 0)$

3. The incidence or hazard function of $T$ can be derived from the PGF as followed:

$$
\lambda(t_0, t) = -\frac{\psi'(1,0; t_0, t)}{\psi(1,0; t_0, t)}.
$$

Since the process $(I(t), Z(t))$ is Markovian, the PGF $\phi(s,t)$ can be shown to satisfy the Ricatti differential equation:

**Theorem A.0.1** Under assumptions (2)-(5) $\phi(s,t)$ satisfies the following differential equation

$$
\frac{\partial}{\partial t} \phi(s,t) = (x_2 - 1)[x_2 \alpha(t) - \beta(t)] + x_2 (x_3 - 1) \mu(t) \frac{\partial}{\partial x_2} \phi(s,t), \quad \phi(s,s) = x_2.
$$

Using filtered Poisson process, we can link these two PGFs: $\phi$ and $\psi$ in the following theorem.

**Theorem A.0.2** Assume that $X(t) \nu(t)$ is finite for all $t \geq 0$ and that the number of mutations that occur during $[t, t + \Delta t]$ from $X(t)$ normal stem cells follows a Poisson distribution with parameter $X(t) \nu(t) \Delta t + o(\Delta t)$, independently, where $X(t)$ is the number of normal cells at time $t$, then under condition (2) (3) (5), $\psi(t)(t_0, t)$ is given by:

$$
\psi(t_0, t) = \exp \int_{t_0}^{t} X(s) \nu(s) [\phi(s, t) - 1] ds
$$
The proofs of the two theorems can be found in Tan’s book (1991). In this general setup, there is no closed form solution to the Ricatti equation. However, in some special cases we can obtain the analytical form.

A.1 Recursive formula for computing survival and hazard function for piecewise constant case

The following formula is taken from Heidenreich’s paper (1997 [11]). Let us assume that lifetime $[0, t]$ is divided into $k$ nonoverlapping intervals $[t_{j-1}, t_j]$ and $t_0 = 0, t_k = t$ such that in each subinterval all the parameters are constant. We use subscript $j$ to denote the parameters valid in the $j$th interval $[t_{j-1}, t_j]$. Let us begin the computational recipes by introducing some notations:

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

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

and $\gamma_i = \alpha_i - \beta_i - \mu_i, \quad i = 1, \cdots, k$.

Then the hazard function and the survival function are given by:

$$\lambda(t) = \sum_{i=1}^{k} \frac{\nu_i}{\alpha_i} \partial_t \log f_i(t_{i-1}, t),$$

$$S(t) = \exp\left(\sum_{i=1}^{k} \frac{\nu_i}{\alpha_i} \log\left(\frac{B_i - A_i}{f_i(t_{i-1}, t)}\right)\right).$$

We can calculate $\partial_t f_i(t_{i-1}, t)$ by means of the following formula:
\[ \partial_t f_i(t_{i-1}, t) = (\exp(B_i(t_{i-1} - t)) - \exp(A_i(t_{i-1} - t))) \begin{cases} A_k \ B_k & \text{if } i = k \\ \frac{\partial}{\partial t} y_i & \text{if } i < k \end{cases} \]

And \( \frac{\partial}{\partial t} y_i \) can be obtained as follows:

\[ \frac{\partial}{\partial t} y_i = \frac{\alpha_{i-1} \ (B_i - A_i)^2 \exp((A_i + B_i)(t_{i-1} - t_i))}{\alpha_i \ (f_i(t_{i-1}, t))^2} \begin{cases} A_k \ B_k & \text{if } i = k \\ \frac{\partial}{\partial t} y_i & \text{if } i < k \end{cases} \]

Function \( f_i \) could be computed recursively from the formulas below:

\[ f_i(u, t) = (y_i - A_i)\exp(B_i(u - t_i)) + (B_i - y_i)\exp(A_i(u - t_i)), \]

\[ y_i = \frac{\alpha_{i-1}}{\alpha_i} \partial_u \log f_i(u, t)|u = t_{i-1}, \]

with initial condition \( y_k = 0 \).