Determination of the historical changes in primary and secondary risk factors for cancer using U.S. public health records

by

Pablo Herrero Jimenez

S.B. Chemical Engineering
MIT, 1995

S.B. Mathematics with Computer Science
MIT, 1995

SUBMITTED TO THE DIVISION OF BIOENGINEERING AND ENVIRONMENTAL HEALTH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Sci.D. IN TOXICOLOGY AND EPIDEMIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2001
[June 2001]

Signature of Author: _____
Division of Bioengineering and Environmental Health
May 2001

Certified by: _____
William G. Thilly
Thesis Supervisor

Accepted by: _____
Ram Sasisekharan
Chairman, Committee on Graduate Students

This doctoral thesis has been examined by a Committee of the Division of Bioengineering and Environmental Health as follows:

Professor James Sherley _____ Chairman

Professor William Thilly _____ 17 May 01 Co-Supervisor

Professor Stephan Morgenthaler _____ Co-Supervisor

Professor David Schauer _____

Dr. Suresh Moolgavkar _____

Determination of the historical changes in primary and secondary risk factors for cancer using U.S. public health records

by
Pablo Herrero Jimenez

Submitted to the Division of Bioengineering and Environmental Health May 2001
in Partial Fulfillment of the Requirements for the Degree of Doctor of Science

ABSTRACT

Overall cancer mortality rates have risen from about 4% of all deaths in the early 20th century to about 25% of all deaths by the end of the century in the United States. To assess any potential hypotheses for this increase required knowledge of the mortality rate changes specific to each form of cancer, and the time points when these rates had changed. For this purpose, population and cancer mortality data of the U.S. were collected and organized to create age-specific mortality rates for each birth decade from the 1800s forward, delineated by the organ of incidence. Concurrently, cancer survival data were collected so as to correct for any effect of improved treatment on historical changes in cancer mortality rates.

To analyze these data, a mathematical model for the three-stage process of carcinogenesis (initiation, promotion, and progression) was developed to estimate for each birth decade cohort the value of the fraction of the cohort at lifetime risk for that cancer, the value of the growth rate of the respective precancerous lesion, and the values for the mutation rates of normal and precancerous cells in the organ of incidence. This methodology permits the analysis of the potential historical effect of new chemical exposures during the last century on cancer mortality rates. These chemical exposures represent potential risk factors that determine the fraction of the population at risk of developing cancer (lifetime, primary risk factor), or that hasten death by cancer by altering either mutation or cell kinetic rates (accelerating, secondary risk factor.)

COLON CANCER: Application of this model on the colon cancer mortality data resulted in the estimate that 42% of the population in the U.S. was at risk for developing colon cancer, independent of gender or race. More importantly, there was no significant historical change in the calculated fraction at risk for birthyear cohorts from 1860 to 1940, suggesting that the primary risk factors for colon cancer are not environmental.

Although direct observation of *in vivo* mutation rates of colonic cells does not yet exist, the calculated rate for the first initiation mutation in the colon was interestingly found to be similar to the mutation rate observed for the hprt locus in human peripheral T-cells (~2.1 x 10^{-7} per cell year) and the spontaneous mutation rate of the hprt locus of human B-cells in culture. The estimate for initiation mutation rates increased no more than two-fold from the birthyear cohort of 1860 to the birthyear cohort of 1940, except for European American females for which calculated initiation mutation rates were historically invariant, but since the accuracy of primary data for mortality rates and survival rates cannot be ascertained, the apparent small differences might admittedly arise from unknown biases. Evaluation of the parameter of the growth rate of precancerous lesions showed no significant historical change on this parameter. Curiously, the calculated doubling rate of these lesions (~0.17-0.21) was found to be similar to the growth rate

of children, suggesting that the required initiation events have the net effect of potentially reactivating pathways involved in child development.

The predominant historical change in the observed mortality rates for colon cancer occurred only at old ages. $\mathbf{f_h}$, the ratio of the number of deaths attributed to colon cancer to the number of all deaths sharing the risk factors of colon cancer, increased historically. This is consistent with the hypothesis that treatment for the connected diseases, which share the same risk factors as colon cancer, has improved more rapidly than treatment for colon cancer. Alternately, the number of underdiagnosed colon cancer deaths may have decreased historically. The increased risk of dying and actually being correctly reported of dying of colon cancer within the elderly population is consistent with the increase in $\mathbf{f_h}$. The conclusion is therefore that the observed changes in colon cancer mortality rates can be predominantly explained by the increase in the lifespan of the American population.

LUNG CANCER: Data from cohorts born in the early to mid 1800s permitted observation of age-specific lung cancer mortality within populations not yet affected by cigarette use, as confirmed by independent nonsmoker lung cancer rates from smaller control studies (Peto et al, 1988, 1992). Likewise, data for birth year cohorts from 1880 forward permitted observation of the age-specific lung cancer mortality in populations with historically documented levels of cigarette use. Based on the mathematical carcinogenesis model, 10% of nonsmokers were estimated to be at lifetime (primary) risk of death by lung cancer, though less than 1% actually died of lung cancer due to competing forms of death. On the other hand, 94% of smokers, essentially all, were estimated to be at lifetime risk of death by lung cancer though less than 10% actually died of lung cancer. The fraction at lifetime risk for all other birthyear cohorts, consisting of mixed populations of smokers and nonsmokers, was found to be a simple linear function of reported cigarette use, independent of gender or race.

The mathematical carcinogenesis model predicted that the marked increase on lung cancer mortality rates among smokers was due to an elevated growth rate of their precancerous lesions. Growth rates of preneoplastic colonies of both genders and ethnic groups were found to be 0.17 and 0.32 doublings per year for nonsmokers and smokers respectively. However, with regard to the rates of events such as genetic alterations necessary for initiation or promotion, the analyses suggest that there were no differences among smokers and nonsmokers of either gender or ethnic group, assuming that the number, but not necessarily the kind, of rare events needed for initiation or promotion was the same for smokers and nonsmokers at risk.

Furthermore, if the growth rate of a precancerous lesion initiated during the period that an individual smoked is reduced from the estimate of 0.32 doublings per year while smoking to 0.17 after smoking cessation, the mathematical model accurately predicted the incidence data reported among former smokers (data from Peto et al, 2000). These results could not be replicated by alternately reducing mutation rates after smoking cessation. These findings support the conclusion that cigarette smoking causes lung cancer by stimulating the growth of preneoplastic lesions in all smokers, and that the data of age-specific lung cancer mortality rates are inconsistent with the widely held but untested assumption that cigarette use increases the rate of genetic change in human lung epithelial cells.

Any data and mathematical tools mentioned herein can be acquired either at the MIT Center for Environmental Health Science (CEHS) database webpage (http://cehs4.mit.edu), or by contacting Pablo Herrero Jimenez at pherrero@alum.mit.edu, or Prof. William Thilly at thilly@mit.edu. The offices of the MIT CEHS are located in Room 16-743, and can be reached at the phone number, (617) 253-6220.

Thesis Supervisor: William G. Thilly
Title: Professor of Toxicology

ACKNOWLEDGMENTS

I would first like to thank my advisor, Professor William Thilly, for giving me the opportunity to work on this project. His scientific rigor has been a constant reminder that yet little is known regarding the link between the environment and the increasing incidence of cancer among humans, teaching me that just because one hears something on TV, or reads it in a newspaper or even in a scientific journal, does not necessarily make it true. His requirement of statistical analysis of all experiments, including asking the simple question of whether observations could occur by simple chance alone, should be commended by all of the scientific society.

I would also like to thank Grethe Thilly, Peter Southam, MS, and Dr. Aoy Tomita-Mitchell for doing the initial work in this project, creating both the initial impetus to put together the data set and to develop the initial mathematical models, without which certainly this project would not have advanced enough for it to have become of interest not only to my advisor, but to myself. I would also like to thank both the individuals who participated in putting together the U.S. mortality data set, whose identities unfortunately are unknown to me as they precede my time in the lab, and also the students from the MIT courses TOX104, TOX205, and BEH205 whose suggestions helped me further improve the methodology used in this project.

The members of my committee have also been extremely helpful in guiding me through the potential complications and implications that my work contains. In particular I would like to thank Dr. Suresh Moolgavkar for inviting me to work alongside himself, Dr. Georg Luebeck and Dr. Bill Hazelton at the Fred Hutchinson Cancer Research Center whose experiences in developing mathematical models of carcinogenesis have allowed me to better understand and improve my own models. Likewise, I would like to thank Prof. Kari Hemminki of the Karolinska Institute at Huddinge, Sweden, for giving me the opportunity to work with the Swedish cancer database, from which I reached a better understanding of the influence of genetic inheritance on cancer incidence/mortality rates.

Additionally, work by Dr. E.E. Furth and Dr. Elena Gostjeva on the *in vivo* cell kinetics of colon and bronchial cells respectively, have helped bring more value to the calculations produced by this project.

There are many other members of the Thilly 'community' who I am certainly grateful for their company while at MIT. Coincidence curiously brought a close friend, Janice Vatland, back to MIT to the same lab no less, giving me someone with whom I could go just a 'little' crazy from working with so many numbers. A heartfelt thank you to everyone else past and present: Amanda, Andrea, Beth-Ann, Brindha, Cindy, Davicia, Enda, Helen, Hilary, Hiroko, Jackie, Joey, Kathy, Klaudyne, Konstantin, Luisa, Paul, Paula, Paulo, Rita, Wei-Ming, Wen, Wendla, Wendy, and Xiao-Cheng, for all of their support.

And I certainly cannot forget the 'kids' of 1E at East Campus whose levity helped me keep my mind occupied on matters other than work. They are still today a constant reminder to me of the value of camaraderie and the value of just hanging out. The three years I had the pleasure to be a GRT to them were among the most rewarding experiences of my life. I can only hope that they learned as much from me as I certainly learned from them.

This thesis is dedicated to my parents and brother who have had to put up with me demonstrate how much I know (or I think I know) about cancer.

This work was supported by a training grant from the National Institute of Environmental Health Sciences (NIEHS).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

APC	Adenomatous Polyposis Coli
EAM	European American Male
EAF	European American Female
FAP	Familial adenomatous polyposis
HNPCC	Hereditary Non-Polyposis Colon Cancer
ICD	International Classification of Diseases
LOH	Loss of Heterozygosity
LOI	Loss of Imprinting
MNU	Methylnitrosourea
mtDNA	Mitochondrial DNA
NEAM	Non-European American Male
NEAF	Non-European American Female

LIST OF SYMBOLS

F	Fraction at primary risk
f	Correction factor for connected forms of death
$r_{\text{lower case letter}}$	Initiation mutation rate
$r_{\text{upper case letter}}$	Promotion mutation rate
τ	Rate of normal cell turnover
α	Division rate of precancerous cell
β	Death rate of precancerous cell
α_c	Division rate of cancerous cell
β_c	Death rate of cancerous cell
ζ	Growth rate of child: age 0 to 1.5

η	Growth rate of child: age 1.5 through puberty
N_a	Number of cells at risk for initiation at age 'a'
N_{max}	Number of cells at risk for initiation in adult
N_{stem}	Number of stem cells in organs
N_{tu}	Number of cells in turnover unit
N_{prec}	Number of cells in precancerous lesion
OBS(h,t)	Observed mortality rate for all individuals
$P_{OBS}(h,t)$	Expected mortality rate in individual at primary risk
S(h,t)	Survival rate
R(h,t)	Accuracy rate of diagnosis at death
h	Birthyear cohort
t	Age at death/incidence (equated with promotion)
a	Age at initiation
n	number of necessary initiation events
m	number of necessary promotion events
κ_h	product of initiation mutation rates, n, and N_{max}
Δ_h	average time between initiation and promotion
t_{max}	age at which OBS*(h,t) reaches a maximum

1. INTRODUCTION

With the advent of new treatments, better personal health care, and an improved sanitation infrastructure, (i.e. antibiotics, insulin, vitamin supplements, clean water systems), mortality rates by infectious diseases and metabolic disorders, as well as mortality rates among infants, consequently dropped throughout the 20th century in the United States. Whereas more than 25% of Americans died of infectious diseases at the beginning of the 20th century, less than 3% did so in the 1990s (U.S. Bureau of the Census, 1900-1936; U.S. Department of Health and Human Services, 1937-1997). Concurrently, the percentage of deaths in the United States caused by cancer has steadily increased from about 4% at the beginning of the 20th century to about 25% in the 1990s (U.S. Bureau of the Census, 1900-1936; U.S. Department of Health and Human Services, 1937-1997).

This increase in U.S. cancer mortality rates has been postulated to be a result of environmental exposures from new industries and technologies of this century. Chemicals in the environment are hypothesized to induce genetic alterations (mutations) of DNA through the formation of chemical-DNA adducts, which consequently are misrepaired or misreplicated by the cell, thereby altering the 'normal' DNA sequence of that cell (Loechler, 1989). Since cancer is a result of the accumulation of genetic alterations in tumor suppressor and/or oncogenes of a cell (Knudson, 1971), this led to the hypothesis that chemical exposures induce the necessary genetic alterations for cancer, and thereby are the direct cause of the observed increase in cancer mortality rates of this century.

However, there is yet little *in vivo* evidence to demonstrate the effect of chemical exposure on human DNA. As direct experimental exposure studies on humans are logically not feasible, researchers rely on either animal studies or cell culture studies to postulate the potential

effects of chemicals on humans. Mutagenicity studies have shown that exposure to a chemical creates a unique mutational spectrum on the studied DNA sequence (Benzer and Freese, 1958; Coulondre and Miller, 1977; Cariello et al, 1990; Keohavong and Thilly, 1992). The determination of the mutational spectrum of a suspected carcinogen can then be compared to the mutational spectrum observed in tumor samples of individuals known to have been exposed to the suspected carcinogen. In the case that these mutational spectra match, it is then argued that this is evidence of a direct etiology of the cancer by the chemical.

For example, there is little doubt that smoking cigarettes increase an individual's risk of developing lung cancer, as suggested by the preponderance of cigarette smokers among lung cancer victims of the 1940s and 50s (Wynder and Graham, 1950; Doll and Hill, 1952; Hammond and Horn, 1958) and subsequent epidemiological studies. Denissenko et al (1996) determined the mutational spectrum of the TP53 gene caused by benzo[a]pyrene, a suspected carcinogen found in cigarette smoke, showing that the induced mutational hotspots were identical to those seen in lung tumor samples taken from smokers. The initial conclusion was therefore that benzo[a]pyrene is a cause of lung cancer in smokers.

However, more recent studies of the mutational spectrum of the TP53 gene among a larger number of smokers and nonsmokers with lung cancer have shown that these same mutational hotspots are observed independent of smoking status (Rodin and Rodin, 2000). Furthermore, the tumor suppressor gene(s) involved in the initiation of lung cancer are not yet known, and the loss of function of the TP53 gene has not been shown to be a necessary event for lung carcinogenesis. Lastly, TP53 does not appear to be a part of the initiation or promotion processes as mutations appear to arise in sectors, but not the totality, of tumors in which they are measured.

Current research (unpublished) by Dr. Xiao-Cheng Li-Sucholeiki, Dr. Luisa Marcelino, Amanda Gruhl, and Hiroko Sudo, of the Thilly lab at the MIT Center for Environmental Health Science has sought to determine *in vivo* mutant fractions in normal lung epithelial cells of smokers and nonsmokers. With emphasis on the mutational hotspots of the TP53 gene, their preliminary results reveal no significant differences in the mutant fraction and thereby mutation rate, of bronchial epithelial tissue samples taken from smokers and nonsmokers of similar age (as of this date, three and two lungs of each respectively). This is insofar consistent with previous research that noted no significant differences in the mutant fraction of mitochondrial DNA of normal bronchial epithelial cells taken from smokers and nonsmokers (Coller et al, 1998).

Additionally, indirect effects of external stimuli must be considered seriously before one can conclude that a chemical induces cancer through the induction of genetic alterations. For example, rats treated with methylnitrosourea (MNU) develop mammary tumors. However, MNU was shown to select previously occurring mutants in the H-ras gene rather than inducing the specific G→A transitions as had been previously assumed (Cha et al, 1994). The selection of cells containing independently generated mutations might occur as a result of chemical contact with the affected cells or a more general effect on all cells by releasing growth signals, by mimicking these growth signals, or by suppressing a body's ability to eliminate precancerous cells.

In the case of human colon cancer, a conclusive cause is yet unknown. The loss of function of the APC gene is associated with the initiation of most cases of sporadic colorectal cancer (Powell et al, 1992). Brindha Muniappan of the Thilly lab has sought to assess the possibility that APC hotspot mutations were induced through DNA misreplication. So far she has identified 6 mutational hotspots induced through *in vitro* replication of APC by DNA polymerase

β (unpublished). Of these 6 hotspot mutations, 3 were found to be concordant with the APC mutational spectrum of cells taken from colon cancer samples (accounting for more than 50% of the APC mutations, by frequency, observed in colon cancer patients). This work suggests that endogenous processes rather than chemical exposures may be involved in the early stage of colon carcinogenesis.

Barring further experimental evidence, it cannot yet be concluded whether or not new chemical exposures during this century have led to the dramatic increase in human cancer mortality in the U.S. The hypotheses behind this increase are that chemical exposures directly induce the necessary DNA mutations for carcinogenesis, permit the selective growth of specific mutant cell populations, or accelerate the onset of a cancer independently induced by endogenous processes. Alternately, the overall shift in U.S. cancer mortality rates could simply be a consequence of the increase in the lifespan of the American population due to the improved treatment and prevention of other formerly prevalent forms of death.

In seeking an alternate method to determine the reasons for the increased risk of developing cancer seen among Americans, all available cancer mortality data in the U.S. were collected starting with the reporting year of 1900. Based on previous analytical models of carcinogenesis (Nordling, 1953; Armitage and Doll, 1954, 1957; Knudson, 1971; Moolgavkar et al, 1979, 1981, 1988, 1990a, 1990b, 1992; Dewanji et al, 1989, 1991), a mathematical model was derived for the three-stage carcinogenesis process of initiation, promotion, and progression, which further accounts for the possibility that only a fraction of each birthyear cohort is at risk of developing cancer, defined by the interaction between inherited traits and environmental risk factors.

Previous analytical models have been used to determine such factors as mutation rates of normal and precancerous cells, the number of such necessary events, and the growth rates of precancerous lesions, but only among individuals of the same birthyear cohort. As an extension, application of the quantitative model to existing U.S. mortality data for multiple birthyear cohorts allows the determination of the historical changes, if any, in mutation rates of normal and precancerous cells, and growth rates of precancerous lesions for individuals who died of cancer. Historical changes in any of these parameters consequently permit inferences regarding the mechanisms by which new environmental exposures may have affected the mortality rates of these cancers in this century. Likewise, the calculation of the fraction at risk for a given cancer as a function of an individual's birthyear can serve as a marker of the changes in the suspected fraction of the population exposed to any necessary environmental risk factor.

Special emphasis has been placed on assessing colon cancer and lung cancer mortality rates. Tested herein is the hypothesis that risk for colon cancer is not environmental, as suggested by the preliminary result that the majority of APC hotspot mutations, by frequency, are induced endogenously; the expectation of the analyses is of a historically invariant fraction at risk. Additionally, tested herein is the hypothesis that the increased risk in developing lung cancer among smokers is consistent with Denissenko et al's (1996) observation that chemicals found in cigarette smoke induce genetic alterations, or alternately that mutation rates in smokers and nonsmokers are similar (Coller et al, 1998; Li-Sucholeiki et al, unpublished), but cells containing previously occurring mutations selectively grow when exposed to cigarette smoke.

2. LITERATURE REVIEW

2.1 MATHEMATICAL MODELING OF CARCINOGENESIS

2.1.1 Mortality Data – Birthyear-specific vs. Cross-sectional

Kermack et al (1934) first reported that mortality trends in England and Wales, Scotland, and Sweden were best described as a function of age times a function of the date of birth. Conceptually, the argument is that environmental exposures throughout life are the determining factors of both the expected age at which death will occur as well as the potential causes of death. Given that individuals born in the 1800s would have experienced considerably different environmental exposures early in their lives than, for example, individuals born in the 1900s, the age-specific function of mortality would be predicted to vary by birthyear cohort.

Were one to instead consider mortality rates for individuals who were alive during, for example, the 1970s (Cross-sectional mortality data), each age group of this cohort consists of individuals born during different historical periods. For example, the existing population of 90-year olds alive in the 1970s would have been born in the 1880s. Smoking prevalence among U.S. women born in the 1880s was less than 3% (Harris, 1983). In contrast, the existing population of 50-years olds alive in the 1970s would have been born in the 1920s. Smoking prevalence among U.S. women born in the 1920s was 40% (Harris, 1983). As there is an increased risk of dying of lung cancer among smokers (Wynder and Graham, 1950; Doll and Hill, 1952; Hammond and Horn, 1958), the average 90-year old and the average 50-year old who were alive in 1970 would have had a different experience in their exposure to cigarette smoke, and thereby a different likelihood of dying of lung cancer during the year 1970, a likelihood which is not solely dependent on differences in age.

Furthermore, female smokers born in the 1880s began smoking on average at age 31, whereas female smokers born in the 1920s began smoking at age 21 (Harris, 1983). The age-specific onset of lung cancer disease due to smoking cigarettes would therefore be expected to occur 10 years later in the life of a female smoker born in the 1880s than in the life of a female smoker born in the 1920s. In other words, one expects the lung cancer mortality rate of a 50-year old female smoker born in the 1880s to be the same as for a 40-year old female smoker born in the 1920s, and so forth. Only by correcting for the differences in smoking prevalence and age of smoking initiation can one truly compare lung cancer mortality rates among different age groups using cross-sectional mortality data.

However, for other forms of cancer, such as colon cancer, one cannot similarly correct for differences of exposure to an environmental risk factor, as the predominant cause of most cancers are not as well known or as well documented. Therefore, as Kermack et al (1934) suggested, it is simply best to consider individuals by their date of birth to reduce the potential confounding factors of a historically variant environmental exposure to an unknown cancer risk factor.

Additionally, Kermack et al (1934) reported the self-evident need to incorporate the probability of still being alive at any age when modeling hazard (death or incidence) functions, since mortality/incidence rates are reported as the ratio of the number of deaths vs. the number of people still alive. Obviously, an individual can only die once. If only a fraction of a cohort is at risk of dying of a particular form of death, one must thereby consider the relative probability that an individual at risk of death is still alive versus that of an individual not at risk, who is therefore at an overall reduced risk of dying.

Fitting mortality data for all forms of death, Kermack et al (1934) ascertained that the age dependent driving function for all deaths followed a simple Makeham-Gompertz relationship (exponentially increasing). This model fit the English and Scottish males' experience, but not the females'. Likewise, the model did not fit the male or female Swedish data. The significance that the rise in mortality as function of age may be exponential was however not explored.

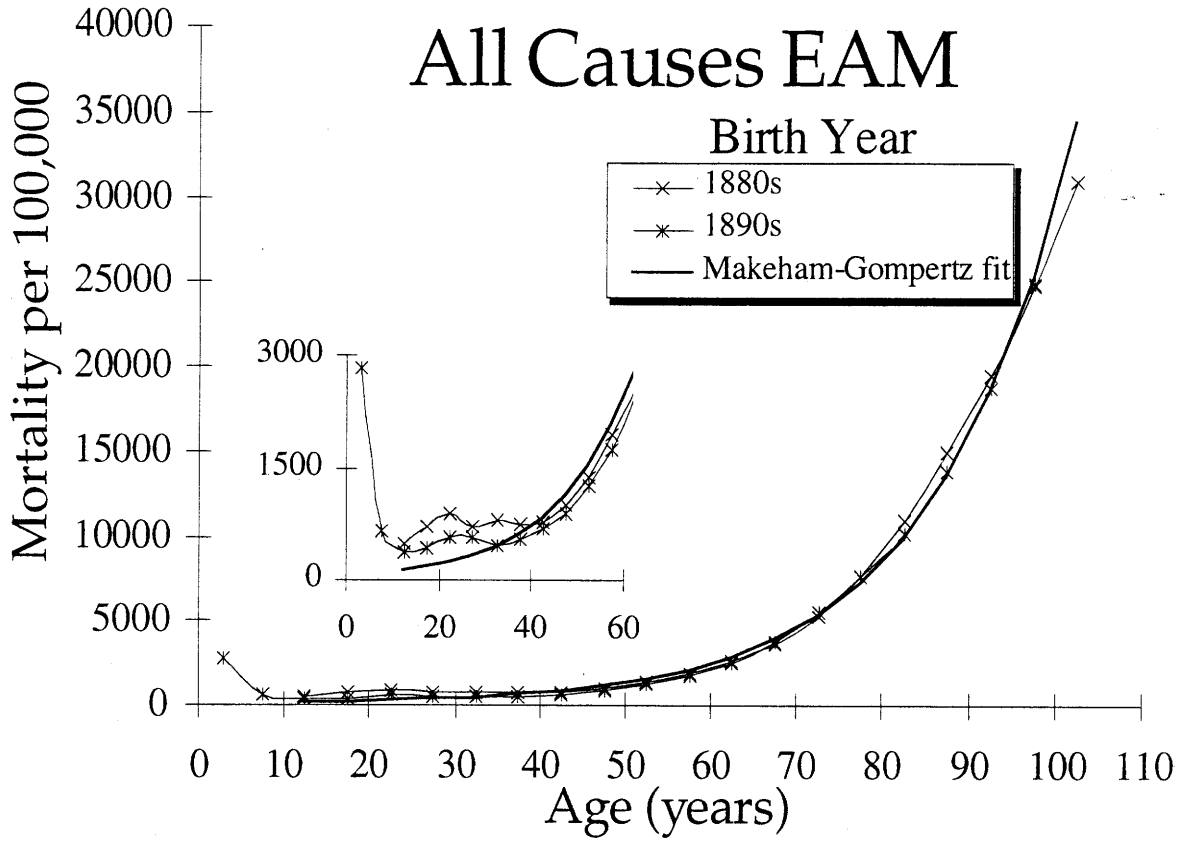
Figure 1 demonstrates the application of the Makeham-Gompertz model to the U.S. mortality data of individuals born in the 1880s, who are more than 40 years of age (See Materials and Methods, Section 3, for description of the collection of data). As was the case for the English and Scottish mortality data among males, the U.S. mortality data does show a Makeham-Gompertz relationship, except for infant mortality rates and mortality rates among 10-40 year olds. These early deaths primarily represent deaths by infectious diseases (as many of these deaths occurred before the discovery of antibiotics), as well as familial forms of cancer, violent forms of death, and other diseases that would predispose an individual to an early death. The early mortality data sets analyzed by Kermack et al (1934) would have also primarily included deaths due to infectious diseases.

However, these forms of death at early age do not necessarily follow a similar age-specific trend as the predominant forms of death among older adults (i.e. cardiovascular disease and cancer). Figure 2 demonstrates that in the U.S., the mortality rates among adults by cardiovascular disease do resemble a Makeham-Gompertz relationship, monotonically increasing for all ages, as had the mortality rates among the elderly from any and all forms of death (Figure 1). However, mortality rates by infectious and parasitic diseases reveal a peak mortality during infancy and a peak mortality throughout adulthood. This illustrates the necessity of analyzing mortality data by each form of death independently.

Fig. 1: Age-specific mortality rate by all forms of death

Mortality rates by all forms of death among European American males (EAM) and European American females (EAF) born in the 1880s and 1890s.

Included is a fit to the 1880s data assuming that the age-specificity of death follows a Makeham-Gompertz relationship as suggested by Kermack et al (1934).


All Causes EAM
Birth Year
1880s
1890s
Makeham-Gompertz fit
Mortality per 100,000
40000
35000
30000
25000
20000
15000
10000
5000
0
3000
1500
0
0
20
40
60
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

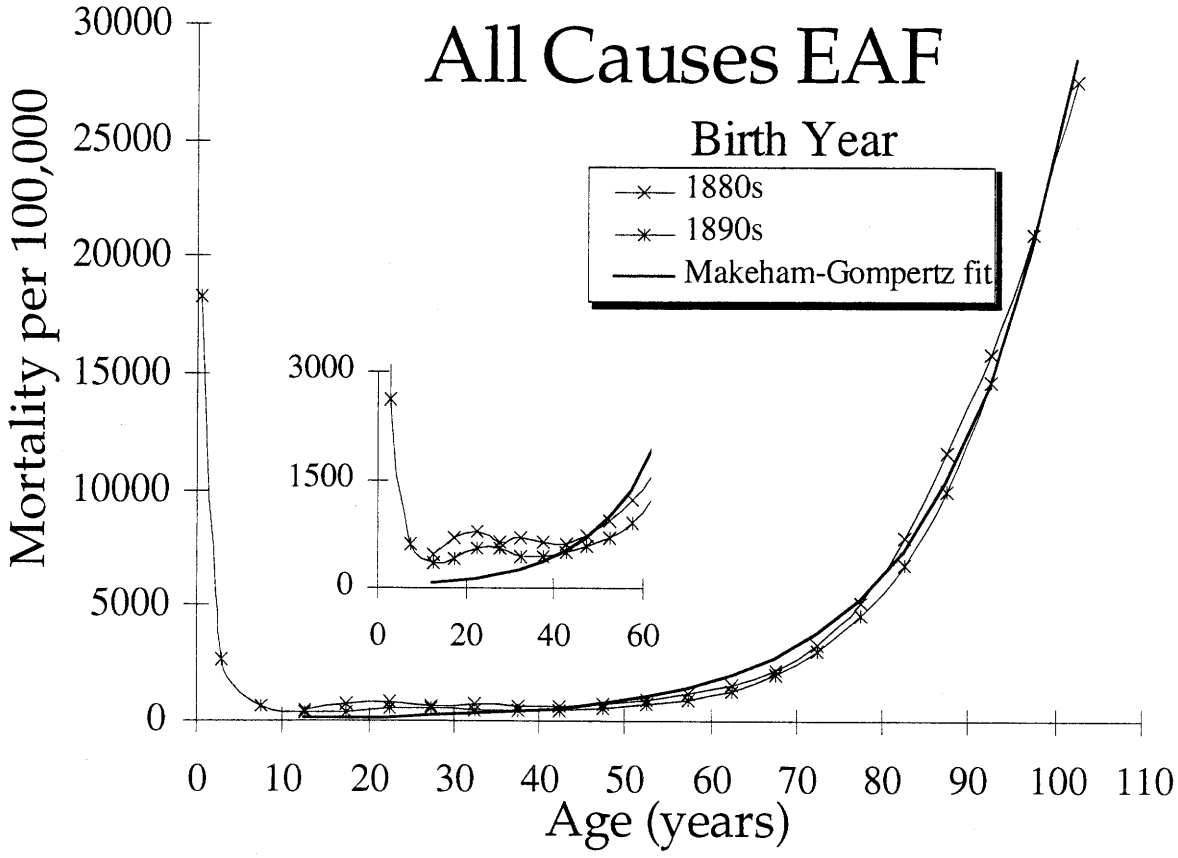
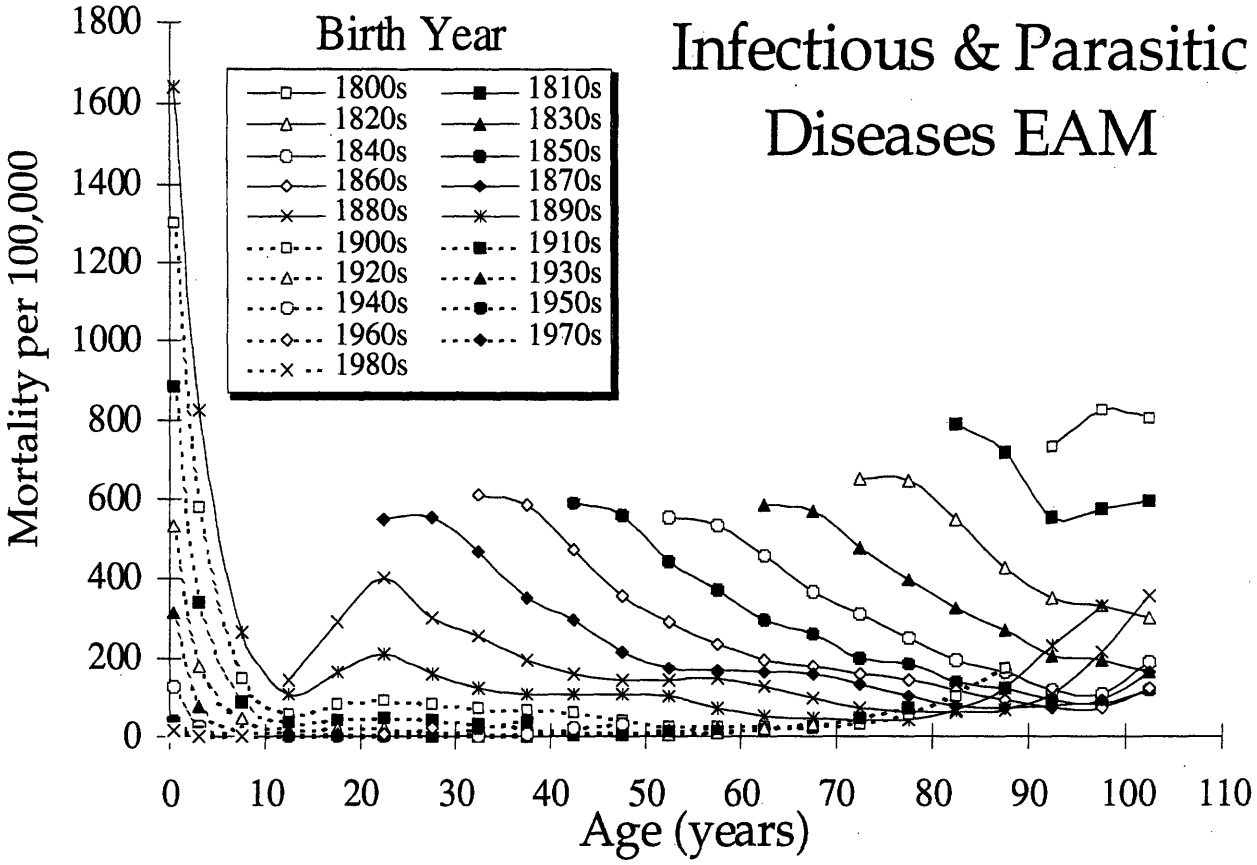

All Causes EAF
Birth Year
1880s
1890s
Makeham-Gompertz fit
Mortality per 100,000
30000
25000
20000
15000
10000
5000
0
3000
1500
0
0
20
40
60
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

Fig. 2: Age-specific mortality rates by infectious diseases and by cardiovascular disease

Mortality rates by infectious/parasitic diseases, and by cardiovascular disease among European American males (EAM) for all birthyear cohorts between the 1800s and the 1980s.


Infectious & Parasitic Diseases EAM
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
Mortality per 100,000
0
200
400
600
800
1000
1200
1400
1600
1800
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

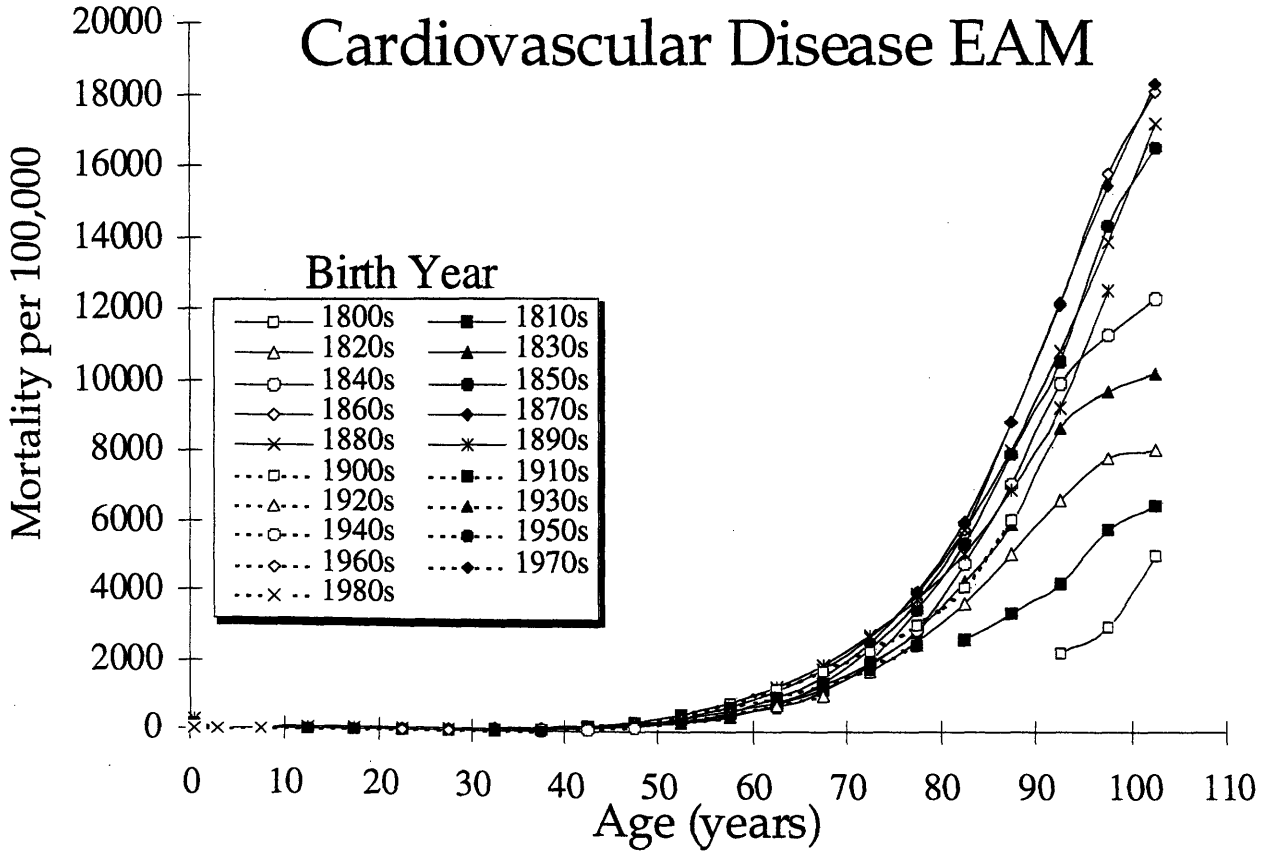

Cardiovascular Disease EAM
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
Mortality per 100,000
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

Figure 2 also reveals that mortality rates by infectious/parasitic diseases has decreased throughout the century, as one would expect with the advent of antibiotics. Curiously one notes that in the more recent birthyear cohorts, there has been an observed increase in mortality by infectious/parasitic diseases among the elderly.

2.1.2 Cancer Mortality Data (Nordling, 1953)

Kennaway (1947) noted that penile cancer development was dependent on the age at circumcision; the earlier the circumcision, the lower the chance of developing cancer. This suggested that unknown events increasing with age could create preneoplastic conditions which were prevented by the act of circumcision. Indeed, mortality rates for all cancers rise monotonically as a function of age, similarly to cardiovascular mortality rates (Figure 2).

Adapting this concept to the general case of all cancers, Nordling (1953) described the first preliminary model of cancer, a process by which a single cell within a cell population of constant size accumulates a series of independent 'changes', now more commonly referred to as genetic alterations, either through mutation or epigenetic effects. After the accumulation of all these events, the cell acquires the capability to create an expanding colony of cells, a tumor. The role envisaged for exogenous agents was to accelerate this chain of events by increasing the likelihood of some, if not all, genetic alterations.

The probability that a cell has acquired the final event necessary for carcinogenesis would directly depend on the probability that there already exists a cell in that individual which has acquired all the other necessary events. Likewise, the probability of acquiring the second to last necessary event for carcinogenesis is itself dependent on the probability that there already exists a cell in that individual which has acquired all but two of the necessary events. This logic

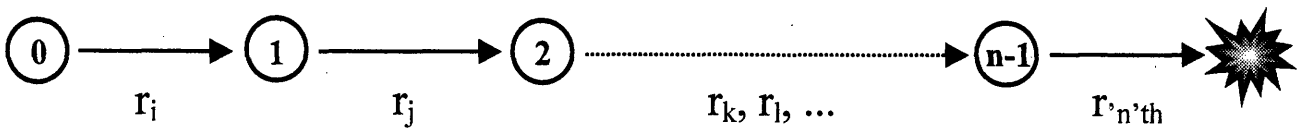
can be carried further to all other previous events. As this allows an individual to acquire cells which have accumulated an incomplete set of necessary events for carcinogenesis, Nordling (1953) deduced that cancer incidence would increase with age as older people are more likely to have cells already containing some but not all necessary events. Quantitatively, mortality could be described as a function of t^{n-1} where 't' is the age at death (or incidence) and 'n' is the number of events required for carcinogenesis (Figure 3). If cardiovascular disease were a process similar to cancer, requiring a certain number of events to occur for an individual to develop the disease, it would explain why the overall mortality rates observed by Kermack et al (1934) among adults appeared to have an exponential (Makeham-Gompertz) relationship.

Nordling concluded that his carcinogenesis model would be incorrect if mortality rates were seen to decrease at higher ages. At the time, some countries did appear to report decreases in mortality by cancer in the older age groups, a trend which he attributed to deaths among older people being designated as "senility" and "other and unknown causes", a trend not seen in the more developed countries. "If, on the other hand, the frequency of cancer actually ceases to increase after a certain age, the hypothesis must be rejected." Nordling did not examine mortality rates past the age of 75 for any country, but did find the best correlation for all cancer data when plotted against time to the sixth power, suggesting 7 mutations were required to get cancer.

Figure 4 represents our own organization of the cancer mortality data in the U.S. as reported since 1900. Cancer mortality rates do in fact rise monotonically to about age 85 as had rates for cardiovascular disease (Figure 2). Contrarily, cancer mortality rates eventually reach a maximum and drop among the elderly. This phenomenon is further explored below in Section 2.1.7. Analysis of our own data set by the Nordling model predicts that 6 events are necessary for carcinogenesis in males, and 5 events are necessary for carcinogenesis in females (Figure 5).

Fig. 3: Nordling (1953) model of carcinogenesis.

Cancer is modeled as the accumulation of 'n' independent events in a single cell, transforming an otherwise normal cell into a cancerous one. The terms r_i, r_j, ... represent the rate of each of these events, where each stage is represented by the number of events already having occurred in the cell. Rate r_i represents the rate at which the first potential event occurs, which can be any one of the 'n' necessary events, assuming that order is not important in carcinogenesis. The 'j'th, or second event, represents any of the remaining 'n-1' necessary events, and so on.



Phenotypically normal cell

\# Number of accumulated mutations

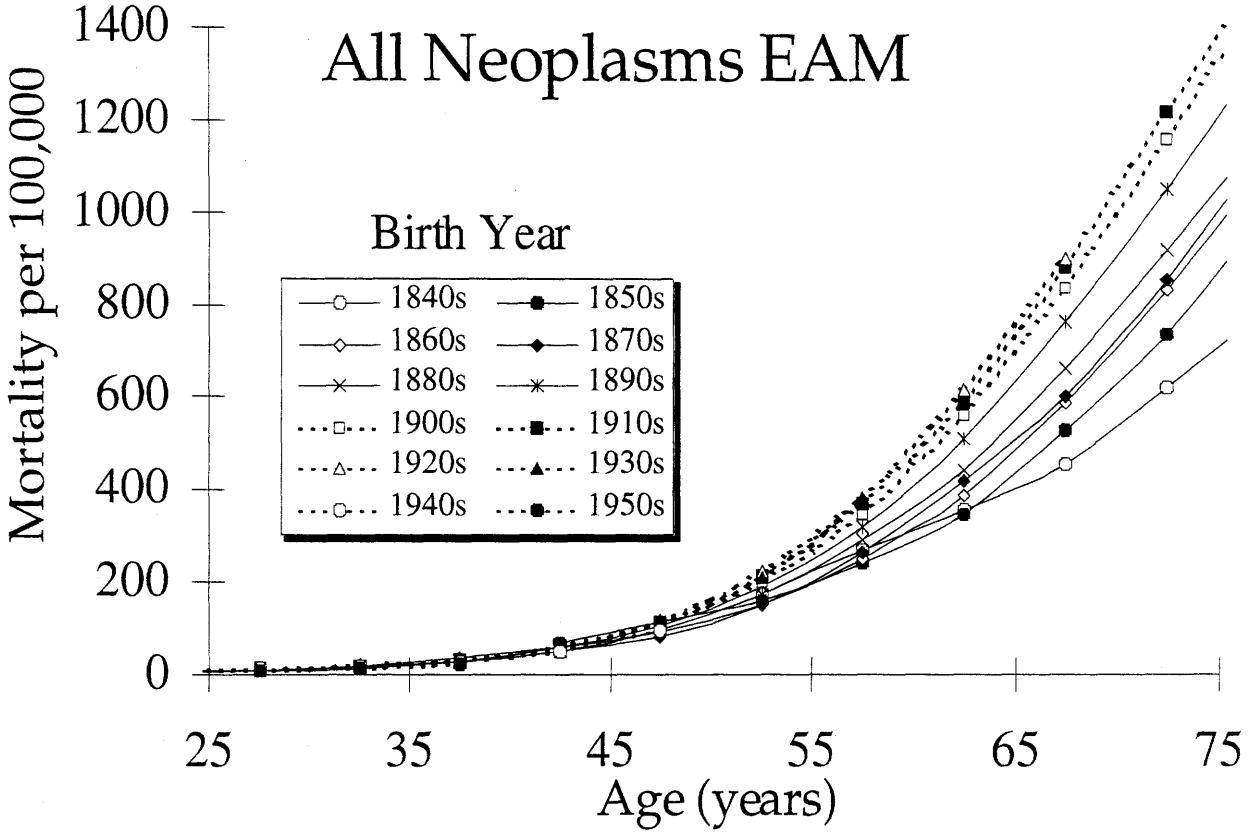
Cancerous cell

Mortality rate as a function of age = Constant x $age^{n-1} = \kappa t^{n-1}$

where the constant κ is proportional to the product of the rates of all 'n' necessary events for carcinogenesis (r_i, r_j, ...) and the number of cells at risk

Fig. 4: Age-specific mortality rate by all cancers

Cancer mortality rates among European American males for birthyear cohorts since the 1800s. If one considers mortality rates for all ages up to 75, birthyear-specific mortality curves appear to monotonically rise for all ages as predicted by the Nordling (1953) carcinogenesis model in which a set of 'n' necessary events must be accumulated by a single cell in order to convert a normal cell into a cancerous one. Inclusion of data past age 75 however reveals that mortality rates not only reach a maximum, but also gradually begin to decrease among the most elderly.


All Neoplasms EAM
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
Mortality per 100,000
1400
1200
1000
800
600
400
200
0
25
35
45
55
65
75
Age (years)

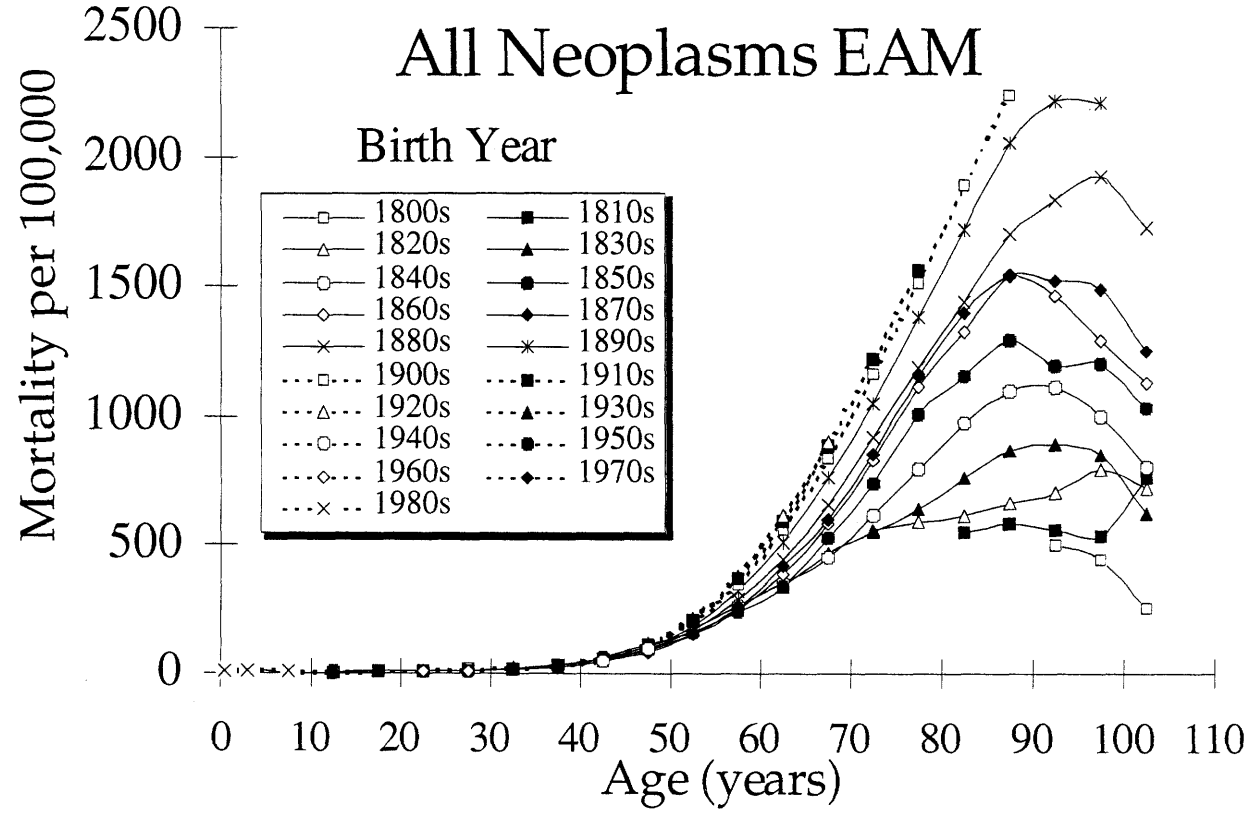
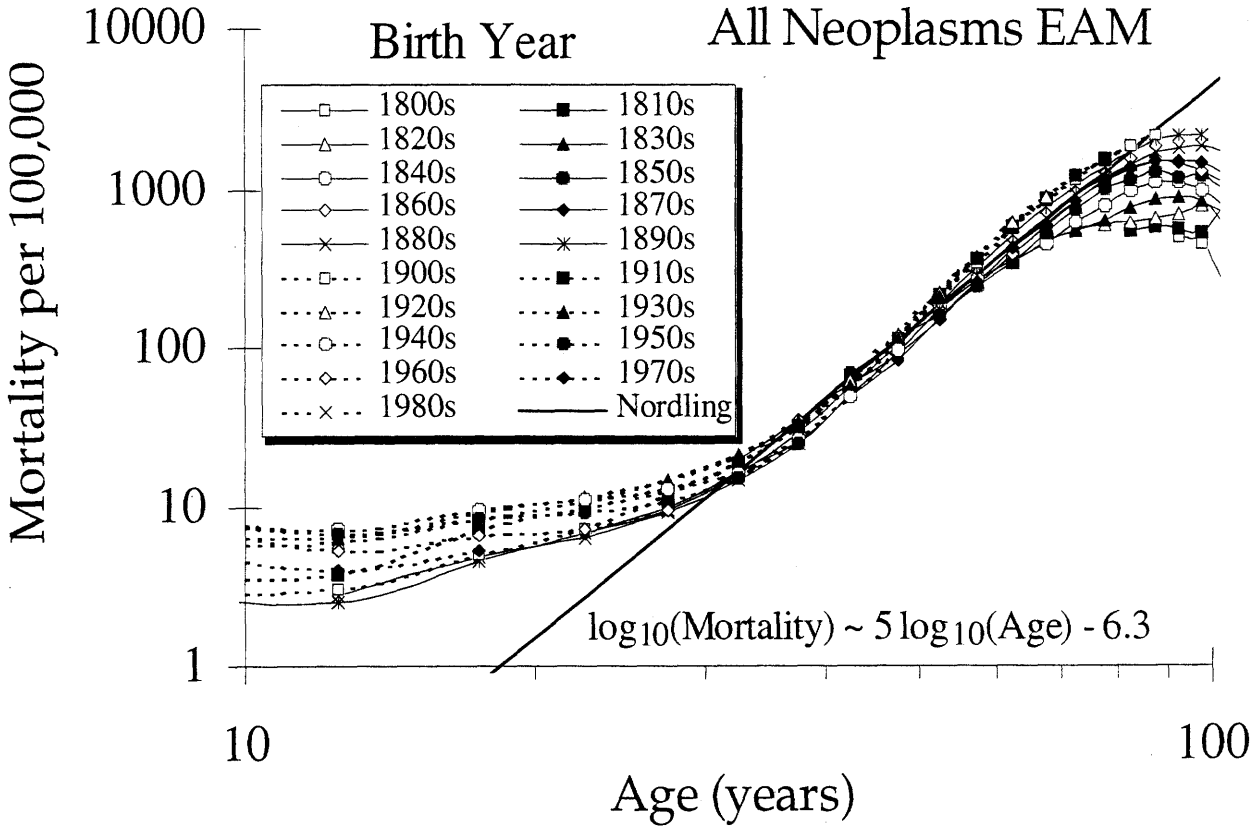
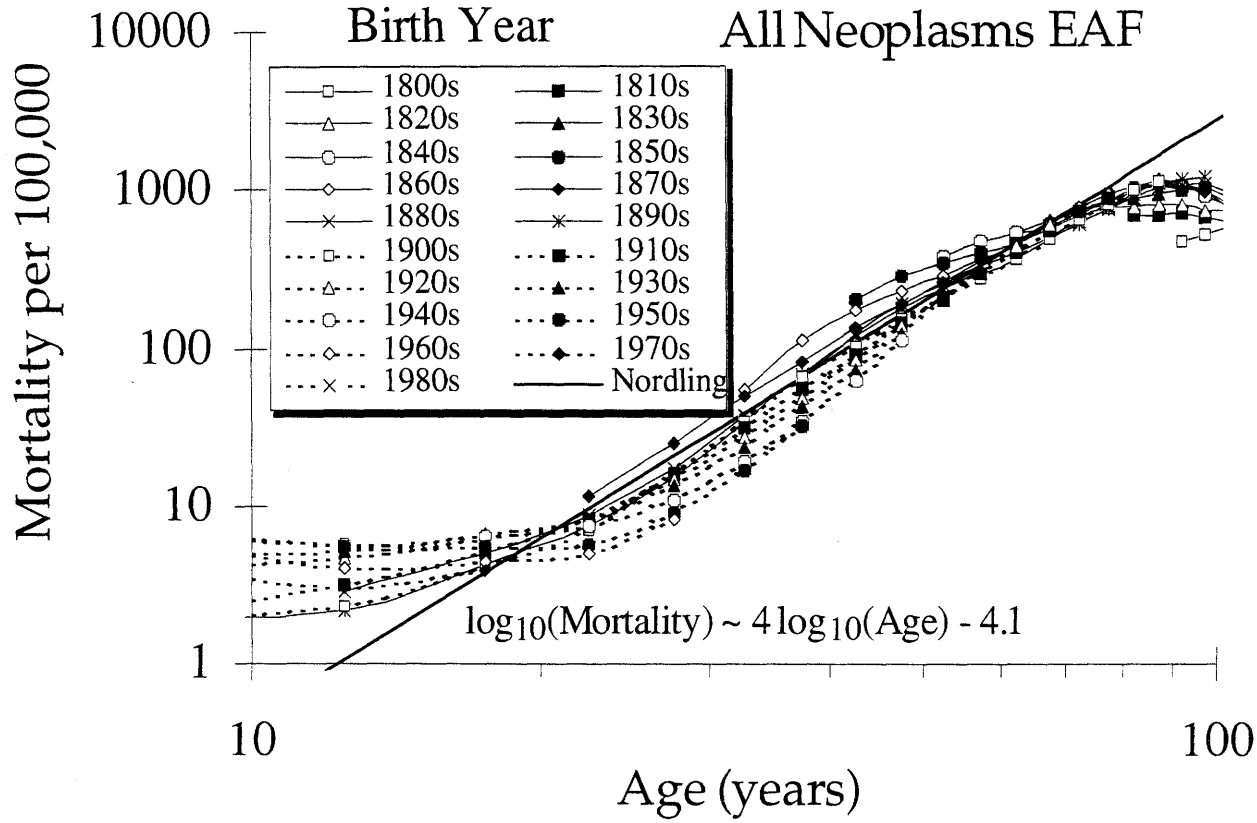

All Neoplasms EAM
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
Mortality per 100,000
2500
2000
1500
1000
500
0
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

Fig. 5: Application of the Nordling (1953) model on U.S. cancer mortality

Cancer mortality rates for European American males (EAM) and females (EAF). The slope of the log-log fit to the cancer mortality rates for ages 35 to 75 predicts that 5 + 1 = 6 events are required for carcinogenesis in males, and 4 + 1 = 5 events are required in females.


All Neoplasms EAM
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
Nordling
log10(Mortality) ~ 5 log10(Age) - 6.3
Mortality per 100,000
10000
1000
100
10
1
10
100
Age (years)


All Neoplasms EAF
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
Nordling
log10(Mortality) ~ 4 log10(Age) - 4.1
Mortality per 100,000
10000
1000
100
10
1
10
100
Age (years)

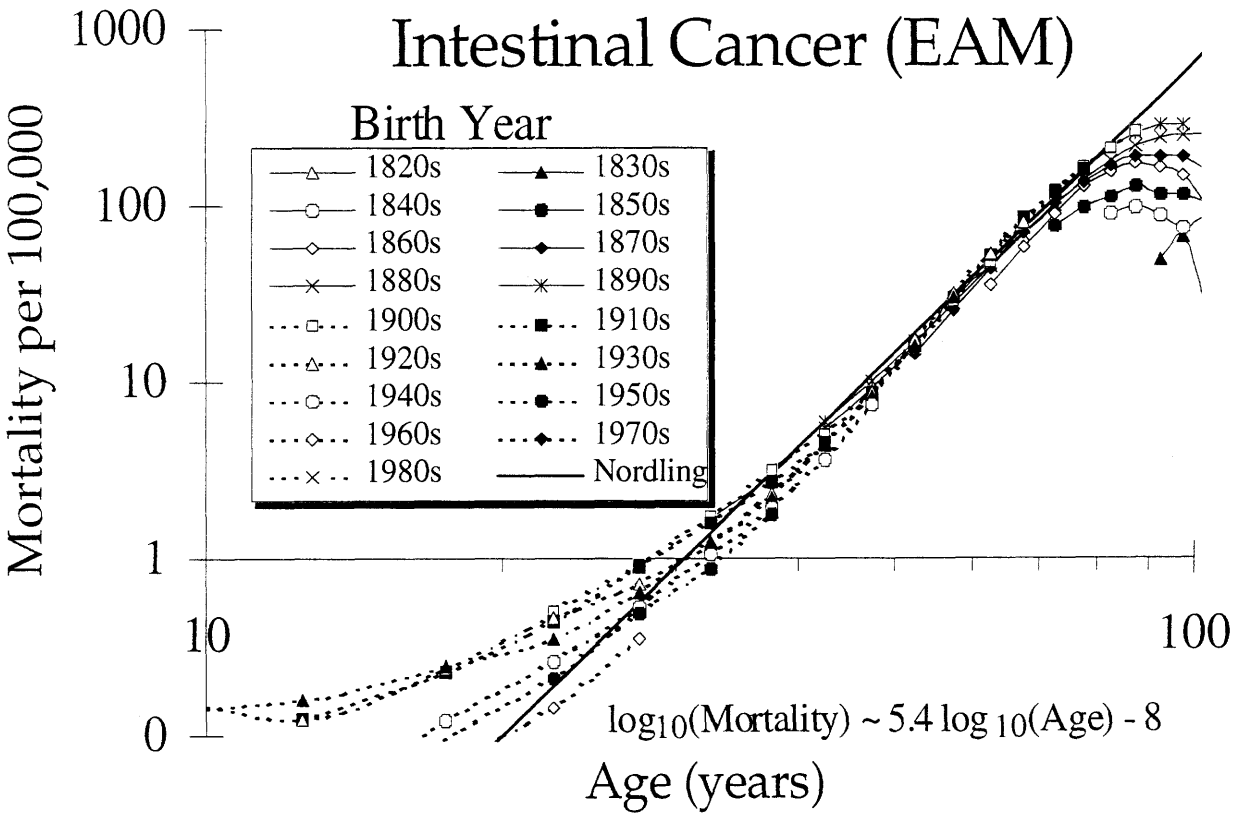
2.1.3 Cancer Mortality Data (Armitage and Doll, 1954)

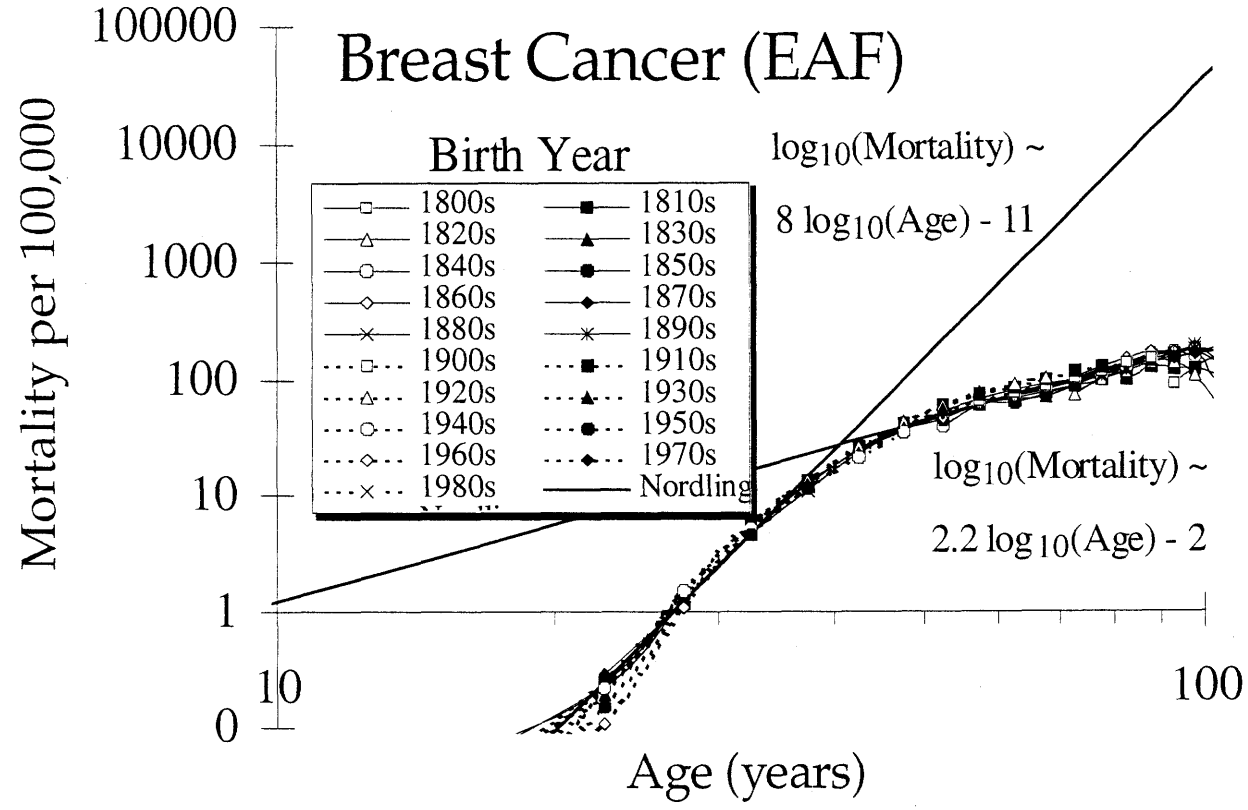
Just as the evaluation of overall mortality rates by Kermack et al (1934) revealed the necessity of evaluating each form of death independently, Armitage and Doll (1954) opted to apply Nordling's hypothesis to individual types of cancers, as well as separating mortality by gender. By doing so, they estimated that the number of mutations required for individual cancers fell within the 6 to 7 range for gastrointestinal cancers, as Nordling had predicted for all cancers. Deviation from the Nordling model below age 30 was attributed to a more susceptible subpopulation that died off early (i.e. people with polyposis coli). Deviation above age 75 was attributed to inaccuracy of the data set. No data above age 85 were apparently available. Furthermore, for gender-specific cancers, such as breast cancer and testicular cancer, Armitage and Doll (1954) found that the Nordling model did not accurately predict the cancer experience assuming any number of necessary events for carcinogenesis.

Figure 6 demonstrates the application of the Nordling model to the U.S. cancer mortality data set for all birthyear cohorts since the 1800s (intestinal cancer among European American males, and breast cancer among European American females). The same conclusions are reached as in Armitage and Doll (1954), of 6 to 7 necessary events for intestinal cancer, and the failure of the Nordling model to accurately predict breast cancer mortality rates, unless it is assumed that two independent subpopulations are at risk for breast cancer. By the Nordling model, 3 times as many events are expected to be required for early breast cancer than for late onset breast cancer, meaning that these events would need to occur more than three time as fast as the 3 to 4 events predicted to be necessary for late onset breast cancer.

Fig. 6: Application of the Nordling (1953) model on intestinal and breast

Intestinal cancer mortality rates among European American males (EAM) and breast cancer mortality rates among European American females (EAF). The slope of the log-log fit to the cancer mortality rates for ages 35 to 75 predicts that 5.4 + 1 = (6 or 7) events are required for intestinal carcinogenesis (Armitage and Doll, 1954). Contrarily, the Nordling model fails to predict breast cancer mortality rates among females, unless breast cancer with early onset represents a subset of the population that dies off faster than the remaining population. In this case, the Nordling model would predict 8 + 1 = 9 necessary events for the early onset form of breast cancer, but 2.2 + 1 = (3 or 4) necessary events for the late onset form of breast cancer.


Intestinal Cancer (EAM)
Birth Year
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
Nordling
Mortality per 100,000
1000
100
10
1
0
10
100
log10(Mortality) ~ 5.4 log 10(Age) - 8
Age (years)


Breast Cancer (EAF)
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
Nordling
log10(Mortality) ~
8 log10(Age) - 11
log10(Mortality) ~
2.2 log 10(Age) - 2
Mortality per 100,000
100000
10000
1000
100
10
1
0
10
100
Age (years)

2.1.4 Precancerous growth (Platt, 1955)

In response to Armitage and Doll's (1954) analysis of cancer mortality curves, Platt (1955) suggested that the capability of a cell to create an expanding colony is not necessarily unique to a cell in a cancerous state, but could theoretically be observed in cells that have accumulated some, but not all of the necessary events of carcinogenesis. These cells would be considered to compose precancerous lesions which like cancerous lesions grow, albeit at a slower rate. Platt further expounded upon the possibility that cigarette smoke increases the kinetic rates of lung epithelial cells. The suggestion permits for two potential explanations as to why cigarette smoke increases the likelihood of developing cancer, either by increasing the turnover rate of normal and/or precancerous lesions, or by increasing the rate of genetic alterations if the time a cell has to repair damaged DNA is diminished by more rapid division.

2.1.5 Cancer Mortality Data (Armitage and Doll, 1957)

At Platt's (1955) suggestion that a cell having undergone a single mutational event may proliferate at a faster rate than a normal cell, Armitage and Doll (1957) reformulated the Nordling model to accommodate two necessary stages, the first creating a proliferating colony (initiation) and the second permitting an inexorable conversion (promotion) to a lethal tumor. They noted that an intermediate, growing colony would not require as high a mutation rate per cell, such that only one event might be needed to reach each stage. However, as with their application of the Nordling model, Armitage and Doll were able to fit the gastrointestinal mortality data well, while other cancers were believed to still be dependent on additional age-dependent factors. They further estimated that the exponential growth rate of precancerous proliferating lesions is about 0.12, which translates to a doubling rate of about 0.18.

Figure 7 summarizes the Armitage and Doll (1957) model for the general case of 'n' necessary initiation events, which create a proliferating precancerous colony, and 'm' subsequent promotion events which create the first cancerous cell. This produces the interesting condition that the precancerous lesion is expected to comprise of a mixed population of growing cells that have acquired all 'n' necessary mutations, but only some of the 'm' necessary promotion mutations. The precancerous lesion will comprise of a growing colony of cells with the 'n' initiation mutations, a growing colony of cells with the 'n' initiation mutations plus one promotion mutation, and so on. This explains the observation of the loss of function of genes, necessary or not, in some but not all cells of lesions. Additionally, the proliferating rate of a precancerous cell may be dependent on the number of promotion events already accumulated.

2.1.6 Cancer Mortality Data (Peto et al, 1975, 1977)

Peto (1977) however pointed out the complications with Nordling's base model. Peto argued that any model, where incidence is proportional to t^{n-1} and 'n' is large, could be fit to mortality data for many different values of 'n', suggesting that there was no easy way to determine which model was most plausible. His recommendation was that knowledge of cell kinetics and *in vivo* carcinogenesis would be required to ascertain the most appropriate model. Peto et al (1975) did however argue that the increase in mortality rates with age is likely due to the accumulation of mutations, rather than an increasing mutation rate with age (Section 2.2.4).

2.1.7 Cancer Mortality Data among the Elderly (Cook et al, 1969)

The apparent maximum in cancer mortality rates in old age was initially recognized but dismissed as an error of diagnosis and/or reporting in the elderly. In an attempt to address the

Fig. 7: Armitage and Doll (1957) model of carcinogenesis

Cancer is modeled as the accumulation of 'n' independent events in a single cell, transforming an otherwise normal cell into a precancerous one which can itself create a proliferating colony (initiation). Within this expanding colony, a cell acquires 'm' further events, creating the first cancerous cell (promotion). The terms r_i, r_j, ... represent the rate of each of the initiation events, while the terms r_A, r_B, ... represent the rate of each of the promotion events. The doubling rate of a cell in the precancerous lesion is represented by the difference between the potential division rate of the cell, α, and the potential death rate of the cell, β. The subscript in $(\alpha - \beta)_X$ represents the last promotion mutation acquired.


Phenotypically normal cell

\# Number of INITIATION events accumulated


INITIATED (precancerous) cell ('n' initiation events)

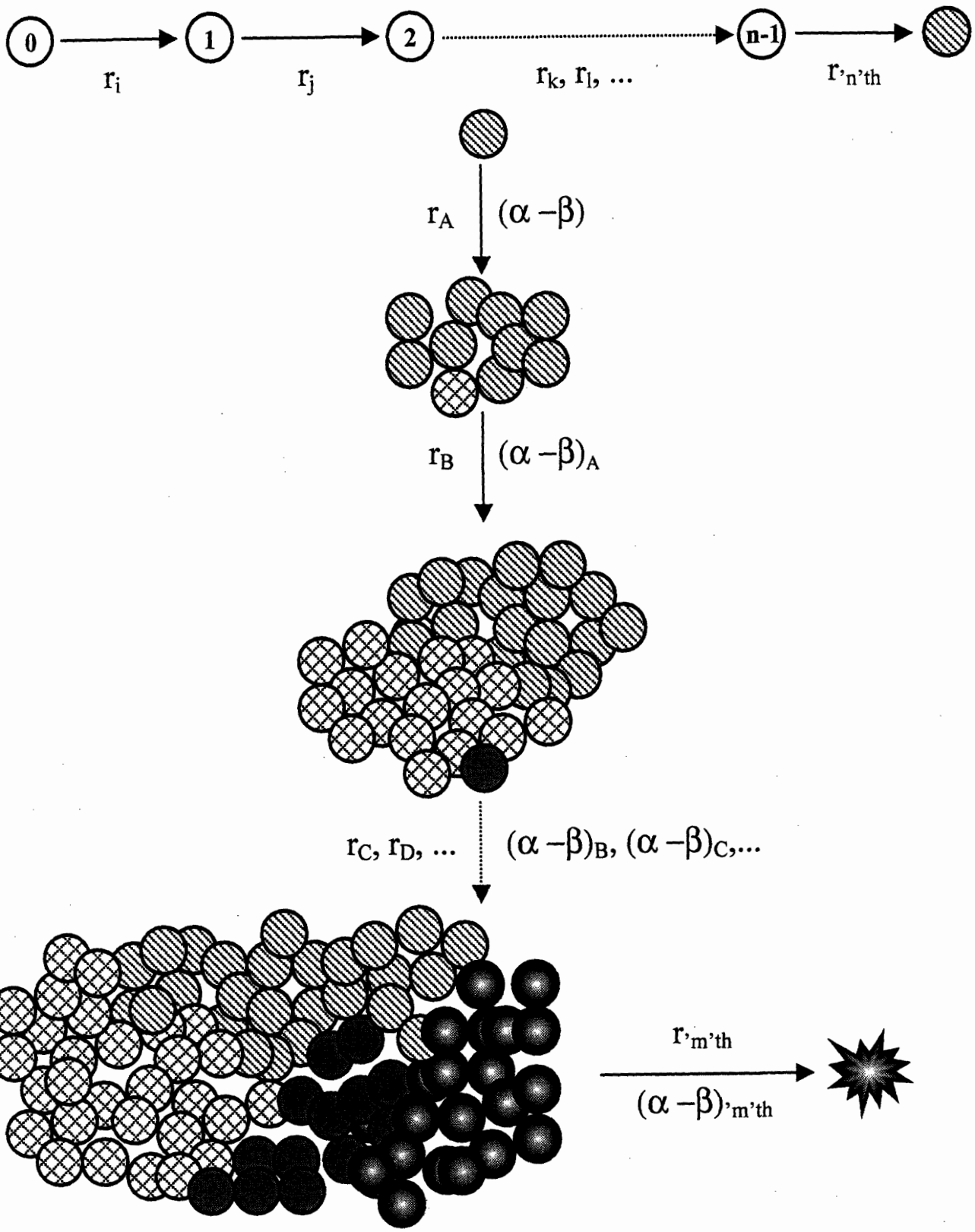

INITIATED cell (acquired 1 promotion event)


INITIATED cell (acquired 2 promotion events)

.....


INITIATED cell (acquired m-1 promotion events)


PROMOTED (cancerous) cell ('n' initiation events + 'm' promotion events)


0
1
2
n-1
r_i
r_j
$r_k, r_l, \ldots$
$r_{'n'th}$
r_A
$(\alpha - \beta)$
r_B
$(\alpha - \beta)_A$
$r_C, r_D, \ldots$
$(\alpha - \beta)_B, (\alpha - \beta)_C, \ldots$
$r_{'m'th}$
$(\alpha - \beta)_{'m'th}$

problem of modeling cancers that did not conform to Nordling's model with or without Platt's modification, Cook et al (1969) were the first to postulate that the recognized downward trend in older patients should not be attributed entirely to misdiagnosis in the elderly. "One explanation of a decreasing rate of increase in old age is that it is an artefact of progressive under-estimation of incidence as age advances...Visual inspection of the graphs show that when curvature is present, it usually occurs throughout the whole range of ages examined, and no relationship was found between the amount of downward curvature and the difficulty of diagnosis, as assessed independently by two colleagues."

They attributed this downward trend to three possibilities. Firstly, they specifically reasoned that a true maximum in the age-dependent mortality and incidence rate would be expected if there were a distinct subpopulation at risk. By virtue of cancer risk, such a subpopulation would have a higher overall death rate than the remaining population which had no risk of cancer. As a birth year cohort aged, there would be a smaller remaining fraction at risk and thus the observed cancer mortality rate in the surviving population could reach a maximum and decline. The mortality and incidence rate of cancer within the subpopulation at risk must therefore eventually decrease back down to zero, as soon as all members at risk have developed that cancer or have died. Cook et al (1969) however does not consider the possibility that these individuals within the subpopulation at risk are not only placed at risk for the cancer of interest, but may also be placed at risk for diseases that share common risk factors with that cancer. This possibility is further explored in Section 3.5.2.2.

Secondly, Cook et al (1969) suggested that incidence could be dependent on the time of first exposure such that the curvature of the mortality curves is representative of the distribution of the age at which an individual was exposed to the necessary environmental risk factor. There

would be a small percentage of the population that is not exposed until late in life and who therefore die at lower rates later in life. It is important to point out that in their analysis Cook et al studied cross-sectional data. For the cohort they studied, individuals living in the 1960s, those who died of lung cancer in their 90's (equivalent to 1870s birthyears) were less likely to have been smokers than those who died in their 70's as smoking prevalence was higher among the later birthyear cohorts (Harris, 1983); as a result, cross-sectional mortality would seem to decrease at the higher ages.

Thirdly, Cook et al (1969) suggested that there may be varying amounts of exposure within the same population, so that people dying at later ages were exposed to less of the carcinogen and thereby develop cancer at a slower rate.

2.1.8. Cancer Mortality Data (Knudson, 1971) and beyond

During the 1970s, Knudson and Moolgavkar, first separately and then in collaboration, made the most extensive attempts to model cancer mortality in terms of the number of required mutations and cell kinetics. Knudson (1971) deduced that children born with a germinal mutation in one of the alleles of the later to be named Rb gene, developed multiple, bilateral retinoblastomas. These were mainly observed in families with retinoblastoma history. Non-familial retinoblastomas were most often detected as single and unilateral. Based on these observations, Knudson formulated a two-mutation model much like Armitage and Doll's (1957), in which the second mutation must occur in a rapidly dividing cell (as retinoblasts are highly proliferative).

Moolgavkar et al (1979, 1981, 1988, 1990a, 1990b, 1992) and Dewanji et al (1989, 1991) further extended this model. They assumed that any cell could either divide and make two new

cells with rate α or could die or terminally differentiate irreversibly at rate β. In the event that this cell had not yet acquired the first (and only) initiation mutation, then $\alpha = \beta$ while in an "initiated" cell, $\alpha > \beta$ An initiated cell would then either become stochastically extinct with probability $(1 - \frac{\alpha - \beta}{\alpha})$ or grow to form a preneoplastic colony in which all cells have an equal probability of experiencing a second mutation and giving rise to a lethal carcinoma. The concept of stochastic extinction permits us to make the prediction that when multiple precancerous lesions are observed in an individual, their average growth rate, $(\alpha - \beta)$, is higher than the average growth rate of a lesion in an individual of the same age who only had a single lesion. Furthermore, the average size of these precancerous lesions is inversely proportional to the growth rate, if one were able to detect the lesions after no more than a few years since initiation.

Moolgavkar's two-mutation model adequately described birth-year dependent but not cross-sectional mortality rates for several cancers, restoring the ideas of Kermack et al (1934) to cancer modeling. Analyses of mortality and incidence data were additionally limited to ages less than 85.

1.2 GENETICS OF CARCINOGENESIS

1.2.1 Initiation of Colon Cancer

Vogelstein's group (1992, 1994a, 1994b) has discovered that loss of both alleles of the tumor suppressor gene, APC, is required for most sporadic colon cancers as well as tumors in familial adenomatous polyposis coli (FAP). (Not all forms of colon tumors exhibit loss of APC, vis. HNPCC patients). Vogelstein proposed that colon carcinogenesis consists of a normal $\rightarrow$ adenoma $\rightarrow$ carcinoma sequence for which loss of both APC alleles is required to obtain an adenoma.

In Moolgavkar's and Knudson's single initiation mutation models, the first APC mutation would have theoretically led to preferential growth of a cell and form an adenoma. However, FAP patients and normal individuals with APC+/- crypts do not exhibit adenomas in such crypts. FAP patients do exhibit increased occurrences of hyperplastic polyps, but these are distinct from adenomas as they are nondysplastic (Jen et al, 1994). To further support the contention that loss of both APC alleles are required to produce a proliferating precancerous colon adenoma, Vogelstein observed that in FAP patients, who are born with a germinal mutation of APC, 80% of their colon adenomas show loss of the inherited wildtype APC copy. After examining the smallest of preneoplastic lesions, 19 of which 24 had mutations in both APC alleles, they concluded that these "... results support the idea that complete inactivation of the APC gene is a critical step for colorectal tumor initiation... Moreover, inactivation was observed in the earliest recognizable phase of tumors, including some lesions as few as two dysplastic crypts." Since their techniques did not detect mutations in the promoter or intron regions and did not detect large deletions, it could not be ascertained whether all adenomas or tumors actually contained a second APC mutation.

Therefore, it was reasonable to suggest that two initiation mutations and at least a single promotion mutation are required to induce neoplastic growth, an assumption justified by the quantitative analyses herein (Section 4.1.1). Moolgavkar and Luebeck (1992) have recognized this point, "Compared with the two-mutation model, the three-mutation model was more consistent with the number of mutations reported to date in colon cancer and with mutation rates measured in the laboratory." Based on their analysis of British colon cancer incidence, they furthermore concluded that there was no growth advantage in a cell containing the first of two APC mutations.

2.2.2 Promotion of Colon Cancer

Little is yet known as to the number of necessary promotion events for colon cancer. Early colon carcinomas and adenomas do demonstrate marked loss of heterozygosity (LOH) for informative loci on all chromosomes, on average 22% (Vogelstein et al, 1988, 1989; Garcia-Patiño et al, 1998; Resta et al, 1998; Uhrhammer et al, 1999; Ragnarsson et al, 1999). Additionally, in a single study, loss of genomic imprinting (LOI) in colorectal carcinomas for a single marker was found to be 44% (Cui et al, 1998). Assuming that promotion entails the loss of heterozygosity or imprinting of 'm' genes, mathematical models could theoretically predict the number of necessary events. For each modeled number 'm', one can estimate an average promotion mutation rate which can then be compared to the equivalent mutation rate that would give rise to the reported levels of LOH/LOI.

2.2.3 Genetic Alterations Observed in Other Cancers

For several other human organs it appears that independent losses of both alleles in certain tumor suppressor genes constitute initiation, i.e. n = 2. The genes and the organs in which their loss appears to represent the only events of initiation are Rb (retinoblastoma, others) VHL (kidney), P16 (melanoma), PTCH(skin), NF1, and NF2 (central nervous system) (Friend et al, 1987; Gnarra et al, 1994; Kamb et al, 1994; Gailani et al, 1996; Rouleau et al, 1993). Others genes involved in initiation of many cancer types are being pursued vigorously by the cancer genetics community. No gene has yet been found to play this role for lung cancer, nor has any gene been found to play a role in promotion for any cancer.

The apparent role of the loss of heterozygosity (LOH) in carcinogenesis is suggested by high LOH levels distributed across the genome in most tumors of epithelial origin. Loss of

heterozygosity by chromatin loss or recombination was not detected in normal bronchial epithelium of nonsmokers (Wistuba et al, 1997; Kohno et al, 1999), while half of 200-800 cell biopsies taken from cigarette smokers were reported to exhibit marked LOH levels (Wistuba et al 1997, 1999a; Kohno et al, 1999). Allelic loss in severely dysplastic colonies averaged 30% among multiple loci. In lung tumors, allelic loss levels vary among chromosomal regions averaging perhaps 60% over all loci analyzed (Wistuba et al, 1997, 1999a, 1999b).

However, we are not able to verify that the reported analytical procedure could account for certain technical difficulties expected in the analysis of small cell isolates in which stochastic sampling errors would become critically important. Whereas Wistuba et al (1997) reported mutant clusters containing as many as 90,000 mutant cells, results from Coller et al (1998) and Li-Sucholeiki et al (unpublished) suggested that mutant clusters are far smaller, 16 to 32 cells. Any conclusion with regard to LOH in bronchial epithelia of smokers and nonsmokers is held in abeyance until these technical concerns are addressed. Additionally, to date, no demonstration has been made that any particular LOH event is required during promotion.

Curiously, many mutations recorded in malignant tumors, such as in the ras proto-oncogenes or the TP53 gene, do not appear to be a part of the initiation or promotion processes as they appear to arise in sectors, but not the totality, of tumors in which they are measured. The absence of these mutations in <u>every</u> carcinoma cell suggests that they are not among the rate-limiting steps in normal tissue cells or precancerous cells which define the age-specific mortality rates. These post-neoplastic mutations may be considered important steps in tumor progression (death from a cancerous tumor) which is considered a relatively rapid process of less than three years duration for most cancers (Axtell et al, 1976).

2.2.4 Somatic Mutation Rates in Normal Tissues

Dr. Aoy Tomita-Mitchell compiled all published reports of the age-specific mutant fraction at the *hprt* locus in human peripheral T cells (Bigbee et al, 1998; Branda et al, 1993; Davies et al, 1992; Finette et al, 1994; Henderson et al, 1986; Hirai et al, 1995; Hou et al, 1995; Huttner et al, 1995; Liu et al, 1997; McGinniss et al, 1990; Tates et al, 1991). Figure 8 shows these mutant fractions as a function of age. These data show a similar distribution around the mean for all age groups, 0-9, 10-19, etc. up to age 75 after which the number of persons with relatively high mutant fractions appears to decline markedly. Using all of the data from ages 0-75, one calculates a constant rate of *hprt* loss of 2.1×10^{-7} mutations per cell year. Additionally, the overall estimated mutation rate for hprt in T-lymphocytes from these studies did not show any changes due to an individual's smoking status, but this could be due to human T-lymphocytes not being the primary target of cigarette smoke mutagenicity *in vivo*.

In human B-cell cultures, the observed spontaneous rates of mutation at the *hprt* locus ranges from 0.5 to 2.5×10^{-7} mutations per cell division (Gennett and Thilly, 1988; Oller and Thilly, 1992; Chen and Thilly, 1996). These *in vivo* and *in vitro* estimates are in reasonable agreement and represent loss of an active gene copy by point mutations and large deletions but not by recombination.

Two estimates of LOH rates in humans differ significantly. Grist et al (1992) reported that the sum of all pathways for loss of heterozygosity of the HLA-A locus in peripheral T-cells was about 6.6×10^{-7} events per cell year. However, Fuller et al (1990) and Jass et al (1994) report observations of colon unicryptal LOH (loss of O-acetylate transferase activity) of about 2×10^{-5} per cell year, 30 times higher than LOH rates seen in blood cells.

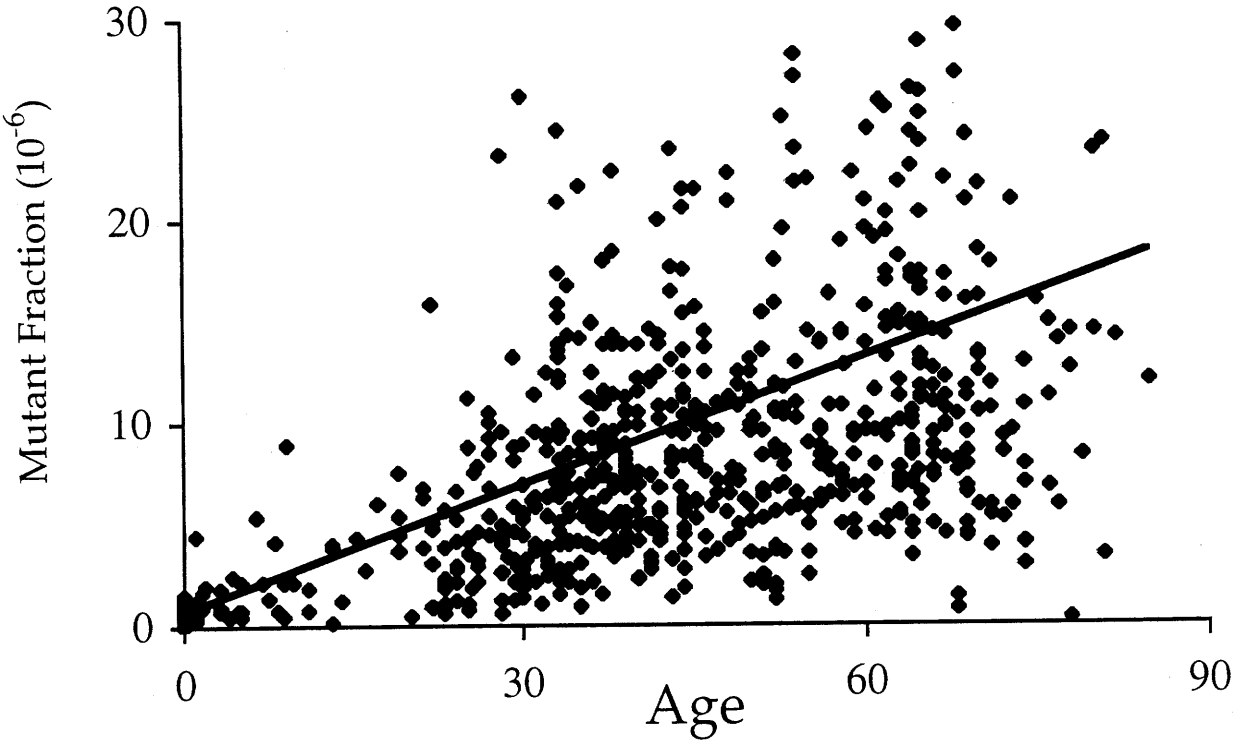
Fig. 8: Mutant fraction of the *hprt* locus of peripheral T-cells

Illustrated as a function of age (n = 740); the slope of the line is 2.1×10^{-7} *hprt* mutations per cell year.

Compiled by Dr. Aoy Tomita-Mitchell

(Bigbee, 1998; Branda, 1993; Davies, 1992; Finette, 1994; Henderson, 1986; Hirai, 1995; Hou, 1995; Huttner, 1995; Liu et al, 1997; McGinniss et al, 1990; Tates et al, 1991)

mutant fraction of the hprt locus of peripheral T-cells



2.3 RISK FACTORS FOR CANCER

2.3.1 Cigarettes - Lung Cancer

The role of cigarette smoke in lung cancer causation was discovered by noting the general rise in lung cancer deaths and the preponderance of cigarette smokers among lung cancer victims in the late 1940s and 1950s. (Wynder and Graham, 1950; Doll and Hill, 1952; Hammond and Horn, 1958) The epidemiological evidence, though clear, did not indicate by what means cigarette smoking acted on lung cells. Cigarette smoke contains small amounts of chemicals such as polycyclic aromatic hydrocarbons, nitrosamines and other substances that mutated cells in culture and also induced tumors in experimental animals. It was thus inferred that cigarette smoking induced tumors by inducing point mutations (Dennisenko et al, 1996; Nesnow et al, 1995; Cloutier et al, 1999; Peterson et al, 1991).

For example, the mutational spectrum of the TP53 gene caused by benzo[a]pyrene included the mutational hotspots seen in lung tumor samples taken from smokers. The initial conclusion was therefore that benzo[a]pyrene, found in cigarette smoke, is a cause of lung cancer in smokers (Denissenko et al, 1996). Rodin and Rodin (2000) have reanalyzed the kinds of nuclear TP53 point mutations occurring in cells of established lung tumors and concluded that these do not differ between smokers and nonsmokers in marked contradistinction to earlier interpretations (Hernandez-Boussard et al, 1998). They concluded that cigarette smoke alternately elevates the risk of lung cancer through the selection of specific mutant cell populations, which are at risk for undergoing further promotion to cancer.

Coller et al (1998) examined the mutational spectra of a 100bp sequence of mitochondrial DNA in bronchial epithelial samples taken from smokers and nonsmokers, including three monozygotic twin pairs, discordant for smoking. Coller et al (1998) concluded that the mtDNA

mutant hotspots found in smokers' samples did not differ from the mtDNA mutant hotspots found in nonsmokers' samplers. Furthermore they did not find a statistically significant elevated mutant fraction among smokers.

Adapting the methodology by Coller et al (1998) to nuclear DNA mutations, Dr. Xiao-Cheng Li-Sucholeiki, Dr. Luisa Marcelino, Amanda Gruhl, and Hiroko Sudo of the Thilly lab at the MIT Center for Environmental Health Science determined *in vivo* mutant frequencies of bronchial epithelial cells of smokers and nonsmokers (unpublished). Given the results by Denissenko et al (1996), they were interested primarily on the hotspots of the TP53 gene. Preliminary results revealed no significant differences in the mutant frequency of bronchial cells of three smokers and two nonsmokers of similar age. The mutant fraction of the hotspot at base pair 746 was found to be on average 4.0×10^{-5} among nonsmokers and 4.7×10^{-5} among smokers, with the highest mutant fraction actually observed in one of the nonsmokers. Alternately, the mutant fraction of the hotspot at base pair 747 was found to be 2.5 times higher in smokers. However, this was not found to be significantly different as the variance of mutant fractions between smokers and nonsmokers was about the same as the variance of mutant fractions within the smoker and within the nonsmoker groups themselves. This leads to the preliminary conclusion that cigarettes elevate the risk of developing lung cancer by means other than a direct mutagenic effect on bronchial DNA.

2.3.2 Methylnitrosourea – Breast Cancer (rats)

When treated with methylnitrosourea (MNU), rats develop mammary tumors. Cha et al (1994) demonstrated that MNU did not actually induce the G→A transitions of the H-ras gene found in all tumor cells of about 85% of these mammary tumors. Alternately, these mutations

were found to arise in preexisting clusters of 100-200 cells. MNU was revealed to select for those cells already containing independently generated mutations. The cause for this selection is yet unknown; possibilities include the induction of the release of normal growth signals in these mutant cells, the mimicking of those signals, or the suppression of internal mechanisms which may help eliminate precancerous cells from the body.

This example in rats provides an alternate hypothesis for how smoking may elevate the risk of developing lung cancer. Cigarette smoke could likewise select for specific mutants in the bronchial epithelium, permitting the proliferation of certain cells that can then undergo promotion. Indeed, Auerbach et al (1957) has described the replacement of the normal lung epithelial architecture due to smoking cigarettes. Smokers' upper bronchi contained frequent areas of epithelial thickenings composed of multiple layers of basal cells, rather than the normal nonsmokers' epithelium which is composed of a layer of basal cells and a layer of ciliated cells. These thickenings may represent the selected growth of metaplastic and/or dysplastic cells. Furthermore, these changes can be effectively reversed by smoking cessation (Auerbach et al, 1962).

Therefore, any potential indirect effect of an external stimulus must be considered seriously before one can conclude that a chemical induces cancer through the induction of genetic alterations.

2.3.3 DNA Misreplication – Colon Cancer?

Unlike lung cancer, a conclusive cause for colon cancer is yet unknown. It has been shown that caloric and dietary intake affects the risk for colon cancer (Giovannucci and Goldin, 1997), but Kinzler and Vogelstein (1996) have suggested that dietary factors might not be

mutagens, but rather behave as irritants that initiate tissue regeneration. An increase in the turnover rate of normal or even precancerous tissue would therefore increase the risk of colon cancer without the need of an exogenous induction of the mutations necessary for colon carcinogenesis.

Research in the Thilly lab by Brindha Muniappan (unpublished) has determined that 3 of the 6 so far identified mutant hotspots of the APC gene induced by DNA polymerase β *in vitro* replication are concordant with the mutant hotspots of the APC of cells taken from colon cancer samples. These 3 mutants comprise more than 50%, by frequency, of all APC mutants observed in colon cancer patients. As initiation of most colon cancer involves the loss of function of the APC gene (Powell et al, 1992), the conclusion is that endogenous processes rather than chemical exposure might be involved in the initiation of colon carcinogenesis.

This does not exclude the possibility of a potential effect of chemical exposure on colon carcinogenesis. Chemical exposures (i.e. diet) may still play a role in smaller subpopulations by accelerating the onset of colon cancer in those individuals, by affecting the error rate of endogenous mechanisms, or by altering cell kinetic rates of preexisting mutant cells. However, these risk factors would not be considered primary, essential risk factors for colon cancer development.

3. MATERIALS AND METHODS

3.1 BIOLOGICAL ASSUMPTIONS

3.1.1 Turnover Unit

Dr. E.E. Furth (University of Pennsylvania Medical School) has observed that in the colon: 1) cells undergoing apoptosis can be found near or at the top of a crypt, or lie within the colonic mucosa where they are being degraded by macrophages, 2) mitoses can be detected all the way along the crypt wall, suggesting that as cells within the crypt divide, new cells are pushed upwards towards the lumen of the colon, replacing those which have been lost through apoptosis, and 3) the stem cell population at the bottom of the crypt helps conserve the overall population of cells within each crypt. It is yet unknown whether each crypt in a human contains a single stem cell or more, but each stem cell is assumed to repopulate an independent portion of the colonic tissue; this portion of a tissue defines a turnover unit.

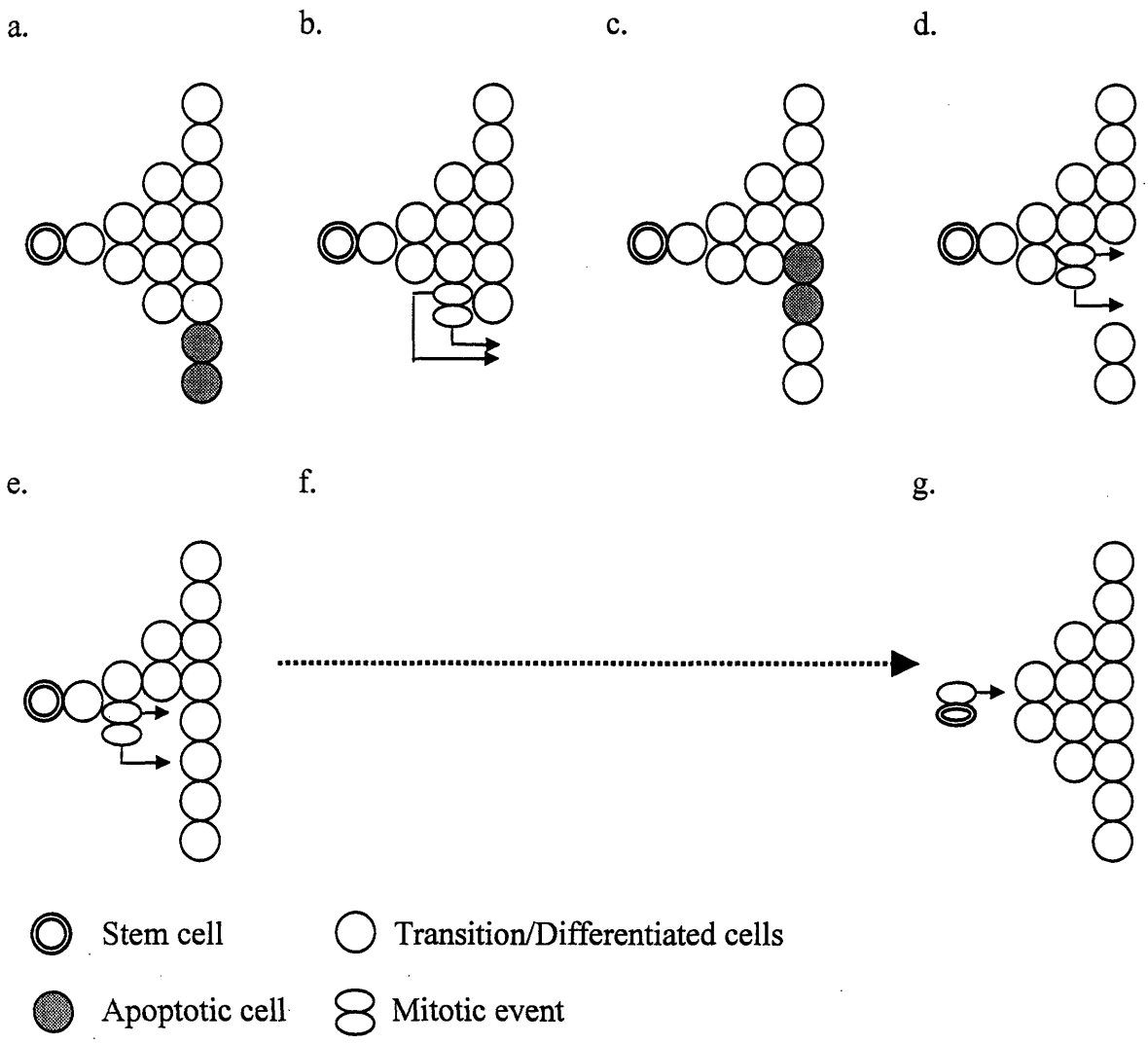
The architecture of the turnover unit in other tissues is not as well understood. In the case of the lung, pluripotent stem cells are assumed to reside in the basal layer of the epithelium, as these comprise the dividing cell population of the normal lung tissue. However, transition cells, non-stem cells which have not yet differentiated into ciliated cells, might also be found in this dividing compartment. As is the case in the colon, stem cells in the lung theoretically regenerate lung tissue by generating transition cells, which themselves can generate other transition cells or fully differentiated cells.

Figure 9 illustrates a basic representation of how a fully-developed turnover unit might undergo tissue renewal. As cells in the terminal layer undergo apoptosis, cells from the previous layers, composed of transition cells, divide and replace those cells which had been lost. As more transition cells divide to replenish the terminal layer, these transition levels themselves must also

Fig. 9: Hypothetical turnover Unit

Illustration of hypothetical model by which a tissue regenerates itself (illustrated for a single turnover unit of the tissue; actual size may vary).

a. Two cells from the terminal layer undergo apoptosis (not necessarily concurrently)
b. A transition cell from the previous layer divides and replaces the two cells lost in step a. (symmetric division)
c. Two more terminal cells undergo apoptosis (not necessarily concurrently)
d. Again, a transition cell from the previous layer divides and replaces lost cells from terminal layer (symmetric division)
e. Since two cells from this last transition layer have now been lost due to migration and differentiation into the terminal layer, a cell from the previous transition layer now divides and replaces these two 'lost' transition cells (symmetric division)
f. As more terminal cells are lost, transition cells continue dividing to replenish both the terminal layer and each transition layer. Eventually, the transition cell which is the direct descendant of a stem cell division itself divides to replenish the second transition layer (symmetric division)
g. Stem cell divides; one cell remains as pluripotent cell; other replaces lost cell in the first transition layer (asymmetric division). This allows for the turnover unit to be renewed while maintaining its stem cell.


a.
b.
c.
d.
e.
f.
g.
Stem cell
Transition/Differentiated cells
Apoptotic cell
Mitotic event

be replenished. Transition cells from previous layers must thus eventually divide to help conserve total cell population within the tissue (Figure 9e). Eventually, the first transition cell, the daughter cell from a stem cell division, must renew the next layer, such that now the stem cell itself must divide to complete the full renewal of the tissue (Figure 9g). The stem cell itself does not migrate to the first transition cell layer; it simply divides to replace the lost first transition cell, thus permanently maintaining a stable cell population in the tissue (asymmetric division). With this final division, this portion of the tissue has undergone full regeneration.

Naturally, tissue may not renew itself in as orderly a fashion as illustrated in Figure 9, but the key point is that the stem cell does not undergo symmetric division. Migration of the stem cell into the transition layer would otherwise lead to loss of the turnover unit, with its eventual differentiation. Furthermore, all other cells would eventually become differentiated and sloughed off through apoptosis.

In this model, transition cells move up one transition layer at a time after each tissue turnover, until it and its daughter cells reach the terminal layer. Within the next turnover, these cells are lost after undergoing apoptosis. If either a transition or terminal cell has acquired one of the necessary mutations required for initiation of a tumor, that mutation would eventually be lost due to normal tissue turnover, unless that particular cell loses the remaining initiation mutations before termination (Figure 10a-e). On the other hand, a mutation acquired by a stem cell would not only never be lost since stem cells do not undergo apoptosis, but also a mutation in the stem cell would eventually lead to the transformation of the previously normal turnover unit into a turnover unit whose all cells contain that mutation (Figure 10f-j).

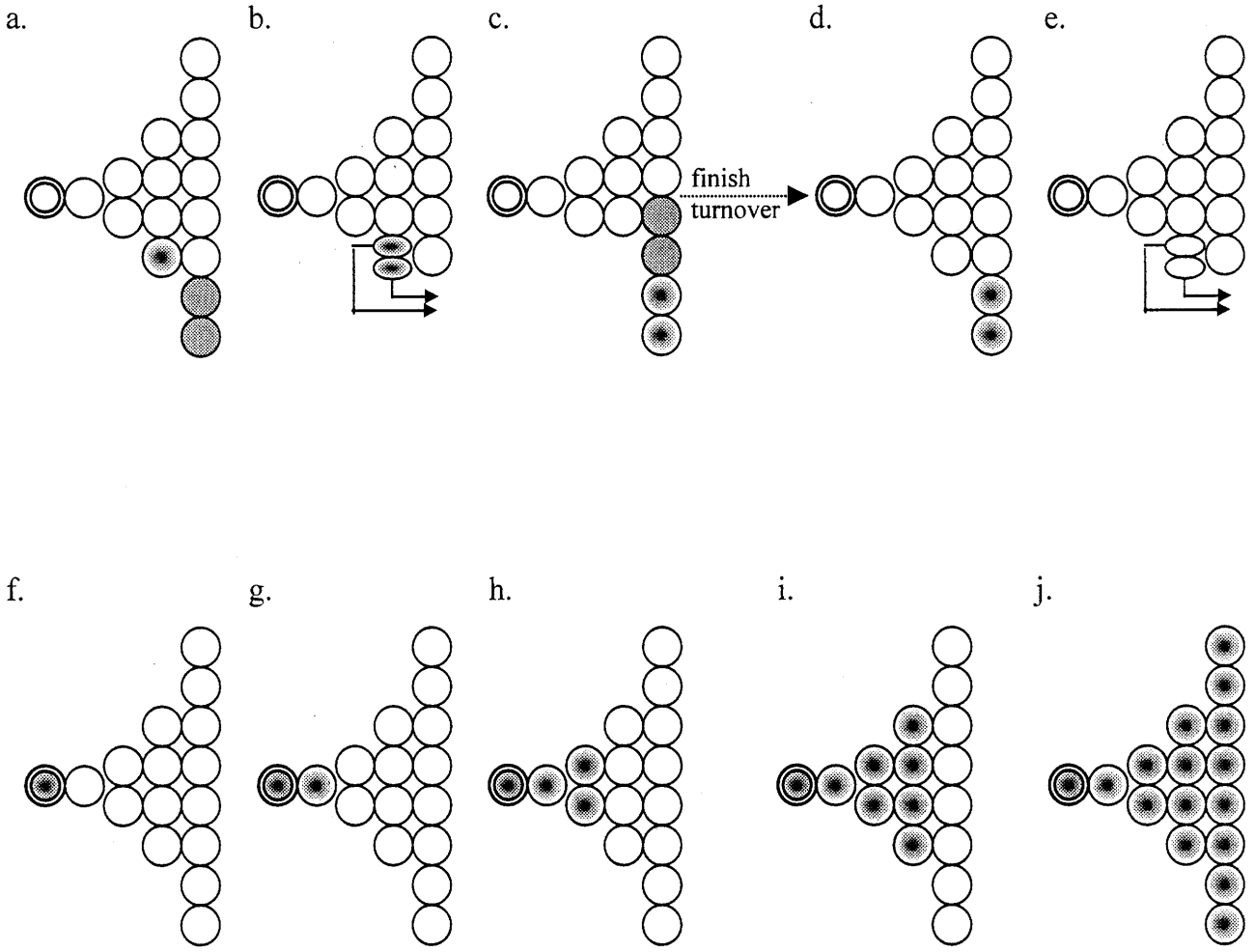
Fig. 10: Redistribution of mutant cells through normal cell turnover

Upper panel; Loss of mutant cells through normal cell turnover

a. Two cells from the terminal layer undergo apoptosis (not necessarily concurrently)
b. A transition cell, which has acquired a mutation, located on the previous layer divides and replaces the two cells lost in step a. (symmetric division)
c. Another two cells from the terminal layer undergo apoptosis (not necessarily concurrently)
d. Rest of turnover unit undergoes further tissue renewal with no more accumulation of single-mutant cells since all other transitional cells are normal
e. Within the next round of turnover, the two mutant cells now in the terminal layer, are replaced by two wildtype cells from previous transition layer. Loss of mutant cells within the turnover unit has occurred. (This occurs regardless of which cell first acquired the mutation, except in the case of the stem cell; See below.)

Lower panel; Transformation of normal turnover unit into a mutant turnover unit through normal cell turnover:

f. Stem cell acquires mutation during a previous division.
g. In the next turnover, the first transition cell (normal) replaces the next layer of transition cells with equally normal cells. Stem cell divides asymmetrically such that the new first transition cell now also carries the same mutation as the stem cell.
h. In the next turnover, the first transition cell (now mutant) replaces the next layer of transition cells with equally mutant cells.
i. Further turnovers subsequently replace transition layers previously normal with cells that carry the same mutation as the stem cell.
j. Eventually, terminal layer is replaced entirely by mutant cells; with a mutant stem cell, this turnover unit cannot reverse to wildtype since the stem cell assures that all transition layers will maintain the mutation found in the stem cell.


a.
b.
c.
d.
e.
finish
turnover
f.
g.
h.
i.
j.
Normal cell
Cell w/ 1st initiation mutation
Apoptotic cell
Mitotic event

If the stem cell has acquired one of the initiation mutations (Figure 10, lower panel), the whole turnover unit is now a target for a second initiation mutation. By the same logic that the first initiation mutation must occur in the stem cell, so must other necessary initiation mutations, as any subsequent mutation in a transition cell or differentiated cell would eventually be lost after normal tissue renewal. The exception is the very last initiation mutation which could theoretically occur in any of the cells. Once a cell has acquired the 'n'th and final initiation mutation, it acquires the capability to produce a proliferating colony, a precancerous lesion, whose kinetics are assumed to be independent of the mechanisms involved in normal cell turnover. All cells in an organ are assumed to be at potential risk for conversion to a precancerous cell. (In the event that a differentiated cell that has accumulated all necessary initiation mutations cannot reversibly continue to divide, then the number of cells at potential risk for initiation is half of all cells in an organ).

Dysplastic colonies of the colon, small adenomas, in fact arise at various positions along the crypt wall, particularly above the area associated with highest mitotic activity, suggesting that the origin of a precancerous lesion is not restricted to the stem cell population, but may include any cells of the transition or terminal layers.

Based on the two initiation mutation model suggested by the genetics of carcinogenesis (Section 2.2), the first initiation mutation would arise in a stem cell, while the second initiation mutation would more likely arise in one of the other cells in the turnover unit. These cells inherited the first mutation as a result of normal tissue renewal within a turnover unit containing a mutant stem cell.

3.1.2 Initiation ('n' events)

As Peto (1977) first pointed out, the first carcinogenesis models (Nordling, 1953; Armitage and Doll, 1954) can accurately fit cancer incidence/mortality data for a large range of number of necessary initiation events, n. It was thereby recommended by Peto that one cannot use mortality curves to determine the number of necessary events, but instead one should rely on *in vivo* observations of carcinogenesis. The following observations can be of use:

1. Small adenomas (precancerous lesions of the colon) already contain loss of normal APC expression. (Powell et al, 1992)
2. FAP patients born heterozygous to the APC mutation have an increased risk of polyp formation; $APC^{+/-}$ crypts are histologically normal (Bjerknes et al, 1997) suggesting that loss of the normal APC allele alone in FAP patients is required for tumor development, but that loss of both APC alleles is needed for tumor development in a normal patient.
3. Mutations in APC result in apoptosis which "...could alter the precise homeostatic balance required in renewing cell populations." (Kinzler and Vogelstein, 1996) Dr. E.E. Furth (University of Pennsylvania Medical School) determined that adenomatous cells show both an elevated rate of cell division and cell death with a slightly higher rate for cell division leading to an overall growth rate for the adenoma; carcinomatous cells show further elevation of cell division rates.

Based on observations 1 and 2, two mutations (loss of APC expression) are required for the initiation of most colon tumors within the non-FAP population.

3.1.3 Growth of Precancerous Lesions + Promotion ('m' events)

Carcinogenesis must include at least one further event (promotion) as suggested by the observation that cancerous cells divide faster than precancerous cells. Figure 11 illustrates our adaptation of Armitage and Doll's multistage model (1957) and the theoretical work of Knudson and Moolgavkar (Knudson, 1971; Moolgavkar et al, 1979, 1981, 1988, 1990a, 1990b, 1992; Dewanji et al, 1989, 1991), using three necessary events for the development of the first cancerous cell (2 initiation + 1 promotion event).

In this model, a stem cell acquires the first initiation mutation in a tumor suppressor gene at a rate r_i. After several rounds of normal turnover (assumed to be completed in negligible time), the turnover unit for that stem cell will be repopulated with single mutant cells, each of which can undergo loss of the second copy of the tumor suppressor gene at a rate r_j. A cell that has undergone both mutations is said to be initiated (precancerous) and now has an elevated division rate, α, and death rate, β.

If the series of divisions and deaths by these precancerous cells is approximately random, then a small precancerous colony could undergo stochastic extinction, as previously described by Moolgavkar (1990b). Assuming that the precancerous colony has grown to a size large enough that extinction no longer is significantly probable (stochastic survival), then the colony undergoes further growth at a deterministic doubling rate of $(\alpha - \beta$. Each of these cells can lastly accumulate a third mutational event at rate r_A (promotion). Accumulation of the promotion event (assuming one) leads to an elevated growth rate of those cells, the initial phenotype of a carcinoma.

Fig. 11: Three-mutation (2 initiation, 1 promotion) model of carcinogenesis

1. A stem cell acquires the first of two initiation mutations at rate r_i.

2. After several rounds of normal cell turnover, stem cell repopulates turnover unit with transition and differentiated cells containing the first initiation mutation.

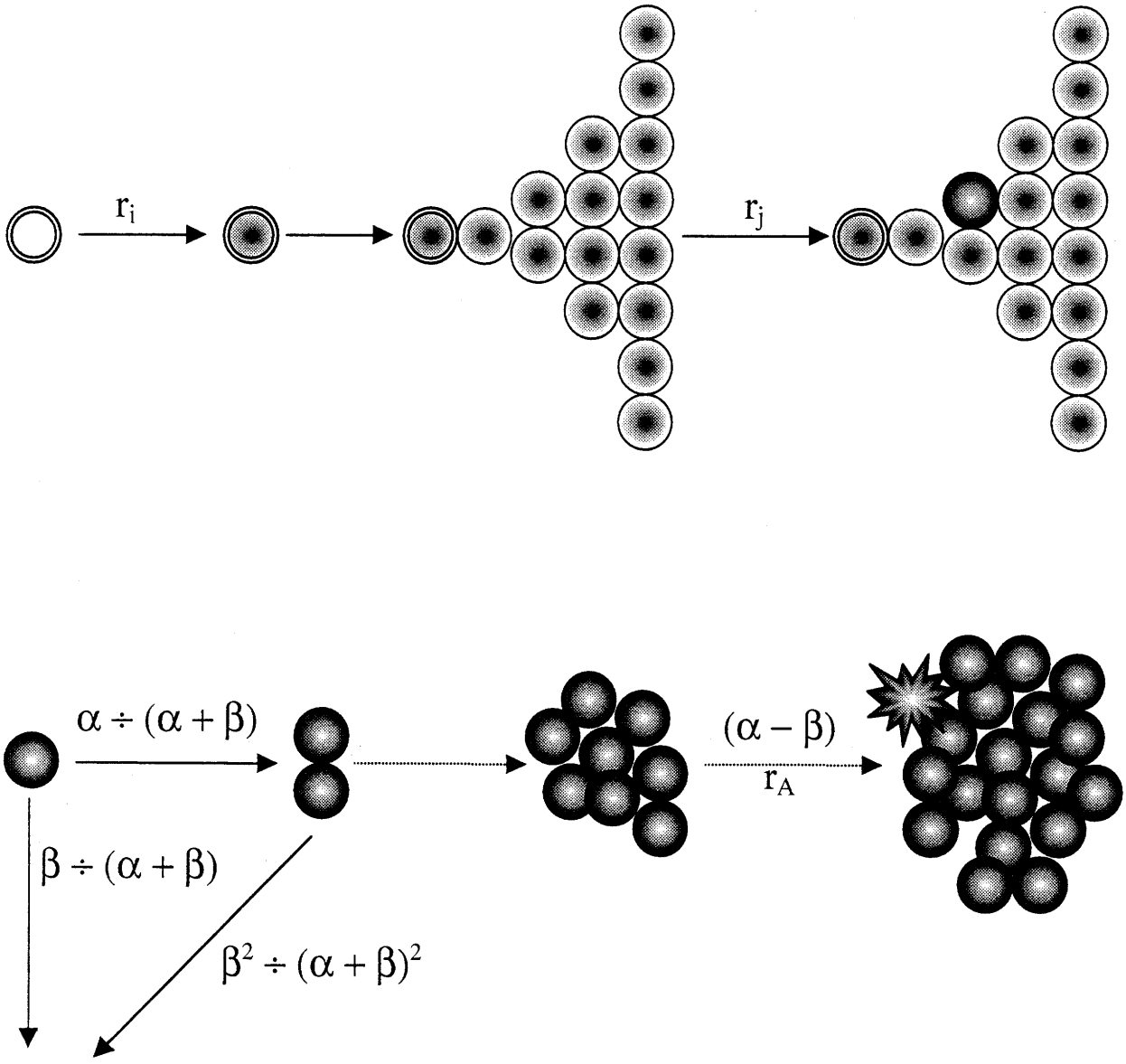
3. One of the cells in the turnover unit acquires the second initiation mutation at rate r_j, becoming an initiated precancerous cell.

2. Initiated cell either undergoes a division with probability $\alpha \div (\alpha+\beta)$ or dies with probability $\beta \div (\alpha+\beta)$, where α represents the average division rate of cells of the precancerous lesion, and β represents the average death rate of cells of the precancerous lesion.

5. Any surviving cell in the precancerous lesion can undergo further division or death by the same process as in 4.

6. If precancerous lesion becomes large enough, complete stochastic extinction becomes significantly improbable, such that the lesion appears to grow at a doubling rate of $(\alpha - \beta)$.

7. Cell in the precancerous lesion acquires a promotion mutation at rate r_A. Assuming only one promotion mutation is needed for carcinogenesis, cell becomes cancerous.


r_i
r_j
$\alpha \div (\alpha + \beta)$
$\beta \div (\alpha + \beta)$
$\beta^2 \div (\alpha + \beta)^2$
$(\alpha - \beta)$
r_A
Stochastic
Extinction
Normal stem cell
Cell w/ 1st initiation mutation
INITIATED (precancerous) cell
PROMOTED (cancerous cell)

3.1.4 Progression

In the three-stage carcinogenesis model, progression consists of a cancerous cell acquiring any further mutations or events needed to lead to the death of the cancer patient (i.e. bringing about metastasis). If subsequent mutations are necessary for death, the elevated growth rate of a carcinoma predictably suggests that these mutations are acquired fast and should not affect the overall model. This third stage occurs rapidly and can be effectively modeled as occurring in zero years. (Axtell et al, 1976)

3.2 PRIMARY DATA SETS

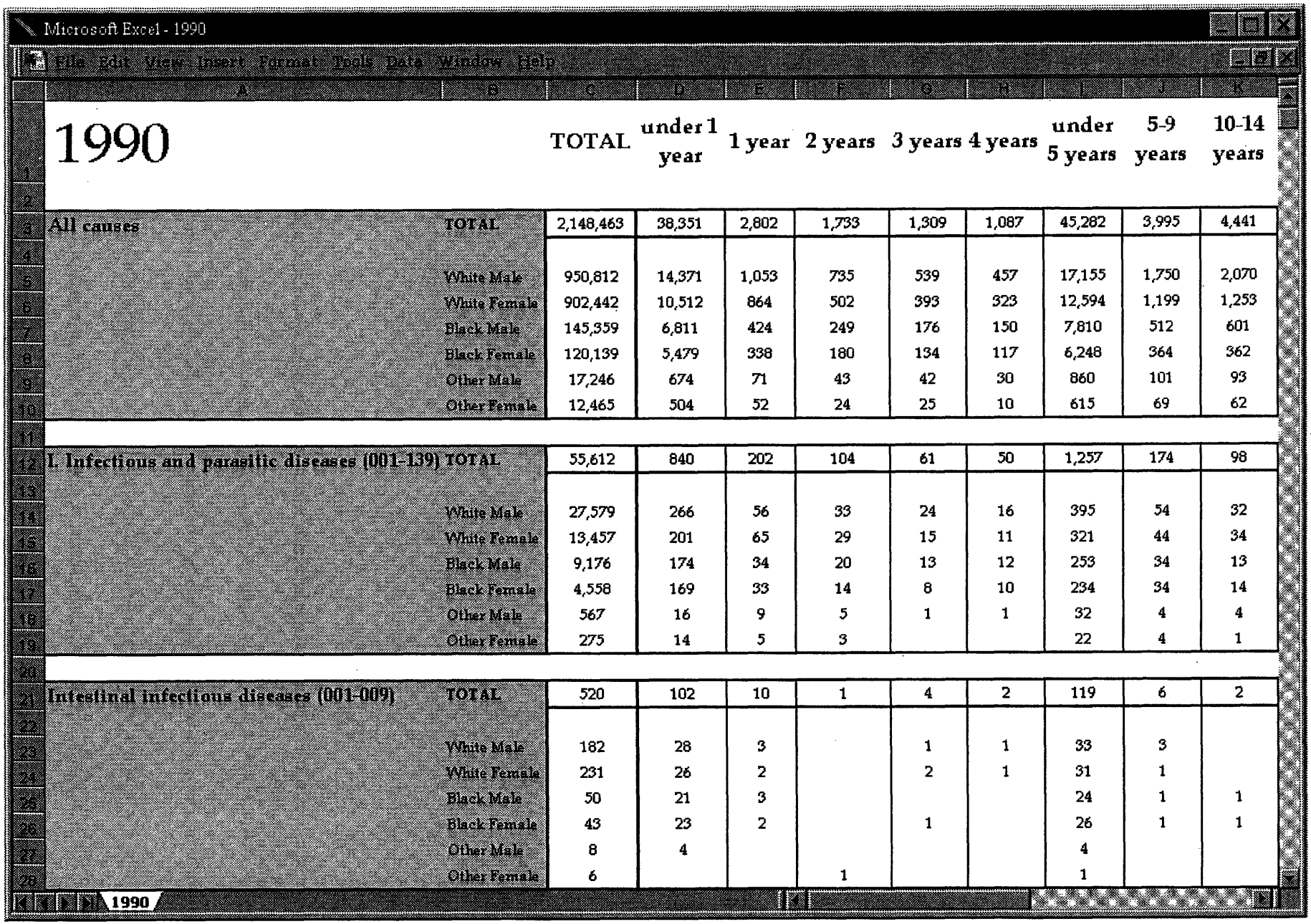
3.2.1. Mortality Data

Annual age-specific mortality data for the U.S. population were obtained from the U.S. Bureau of the Census (Mortality Statistics, 1900-1936) and the U.S. Department of Health and Human Services (Vital Statistics of the United States, 1937-1992). Care was taken to maintain computer records using the same structure as the original sources.

For each reporting year, we created a Microsoft Excel™ spreadsheet listing the year of death, the cause of death, its corresponding International Classification of Disease number, and the number of deaths broken down by age, gender, and race. Figure 12 illustrates an example of one of these spreadsheets for the reporting year 1990. This is an exact replica of the original source, representing our raw data set. The data for White Male/Female represents mortality among European American males/females (including Hispanics), and the sum of the data for Black Male/Female and Other Male/Female (primarily of Asian descent) represents mortality among Non-European American males/females.

Figure 12: Mortality spreadsheet for the reporting year of 1990 in the U.S.

Reproduced from (U.S. Department of Health and Human Services, Vital Statistics of the United States, 1990).



Microsoft Excel - 1990

File Edit View Insert Format Tools Data Window Help

1990		TOTAL	under 1 year	1 year	2 years	3 years	4 years	under 5 years	5-9 years	10-14 years
All causes	TOTAL	2,148,463	38,351	2,802	1,733	1,309	1,087	45,282	3,995	4,441
	White Male	950,812	14,371	1,053	735	539	457	17,155	1,750	2,070
	White Female	902,442	10,512	864	502	393	323	12,594	1,199	1,253
	Black Male	145,359	6,811	424	249	176	150	7,810	512	601
	Black Female	120,139	5,479	338	180	134	117	6,248	364	362
	Other Male	17,246	674	71	43	42	30	860	101	93
	Other Female	12,465	504	52	24	25	10	615	69	62
I. Infectious and parasitic diseases (001-139)	TOTAL	55,612	840	202	104	61	50	1,257	174	98
	White Male	27,579	266	56	33	24	16	395	54	32
	White Female	13,457	201	65	29	15	11	321	44	34
	Black Male	9,176	174	34	20	13	12	253	34	13
	Black Female	4,558	169	33	14	8	10	234	34	14
	Other Male	567	16	9	5	1	1	32	4	4
	Other Female	275	14	5	3			22	4	1
Intestinal infectious diseases (001-009)	TOTAL	520	102	10	1	4	2	119	6	2
	White Male	182	28	3		1	1	33	3	
	White Female	231	26	2		2	1	31	1	
	Black Male	50	21	3				24	1	1
	Black Female	43	23	2		1		26	1	1
	Other Male	8	4					4		
	Other Female	6			1			1		

1990

Figure 13: Mortality spreadsheet for the reporting year of 1990 in the U.S. w/ formulae

Reproduced from (U.S. Department of Health and Human Services, Vital Statistics of the United States, 1990)

Shows incorporated methodology for quality assurance of the raw data set. One can compare calculated vertical and horizontal row totals with reported totals.

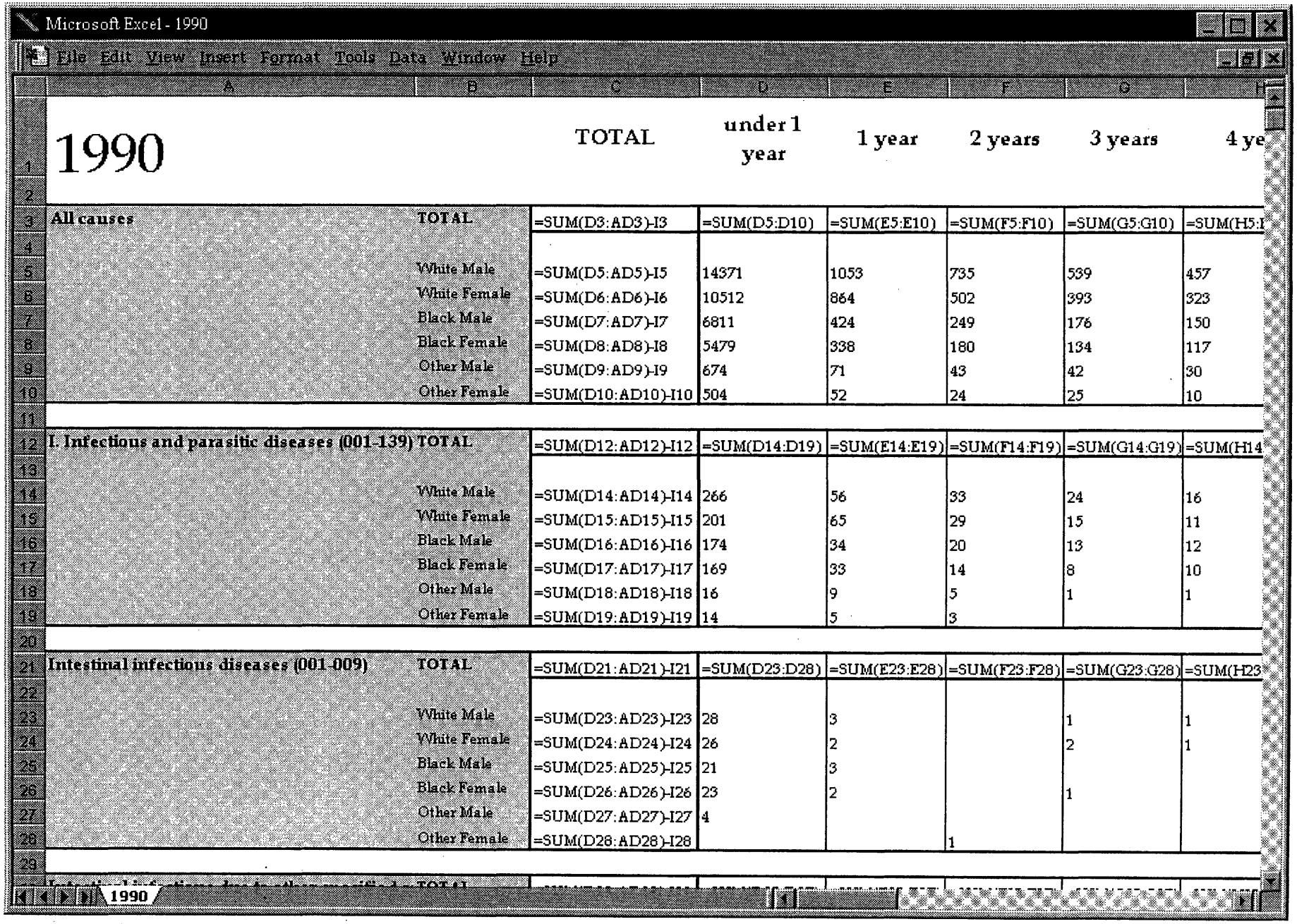
For example:

Cell C5: = SUM(D5:AD5)-I5

SUM of all deaths for White Males (Row 5) for all ages from under 1 year (Column D) to 100+ years (Column AC, not shown) + deaths for individuals of unknown age (Column AD, not shown)

Cell I5 (as can be confirmed in Figure 12) is subtracted as it represents the sum of deaths for under 5 years, which were already included when summing cells D5 through H5.

Alternately, Cell C5 could have been set to =SUM(I5:AD5)



Microsoft Excel - 1990

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F	G	H
1	1990		TOTAL	under 1 year	1 year	2 years	3 years	4 ye
2								
3	All causes	TOTAL	=SUM(D3:AD3)-I3	=SUM(D5:D10)	=SUM(E5:E10)	=SUM(F5:F10)	=SUM(G5:G10)	=SUM(H5:I
4								
5		White Male	=SUM(D5:AD5)-I5	14371	1053	735	539	457
6		White Female	=SUM(D6:AD6)-I6	10512	864	502	393	323
7		Black Male	=SUM(D7:AD7)-I7	6811	424	249	176	150
8		Black Female	=SUM(D8:AD8)-I8	5479	338	180	134	117
9		Other Male	=SUM(D9:AD9)-I9	674	71	43	42	30
10		Other Female	=SUM(D10:AD10)-I10	504	52	24	25	10
11								
12	I. Infectious and parasitic diseases (001-139)	TOTAL	=SUM(D12:AD12)-I12	=SUM(D14:D19)	=SUM(E14:E19)	=SUM(F14:F19)	=SUM(G14:G19)	=SUM(H14
13								
14		White Male	=SUM(D14:AD14)-I14	266	56	33	24	16
15		White Female	=SUM(D15:AD15)-I15	201	65	29	15	11
16		Black Male	=SUM(D16:AD16)-I16	174	34	20	13	12
17		Black Female	=SUM(D17:AD17)-I17	169	33	14	8	10
18		Other Male	=SUM(D18:AD18)-I18	16	9	5	1	1
19		Other Female	=SUM(D19:AD19)-I19	14	5	3		
20								
21	Intestinal infectious diseases (001-009)	TOTAL	=SUM(D21:AD21)-I21	=SUM(D23:D28)	=SUM(E23:E28)	=SUM(F23:F28)	=SUM(G23:G28)	=SUM(H23
22								
23		White Male	=SUM(D23:AD23)-I23	28	3		1	1
24		White Female	=SUM(D24:AD24)-I24	26	2		2	1
25		Black Male	=SUM(D25:AD25)-I25	21	3			
26		Black Female	=SUM(D26:AD26)-I26	23	2		1	
27		Other Male	=SUM(D27:AD27)-I27	4				
28		Other Female	=SUM(D28:AD28)-I28			1		
29								

1990

Ages are reported by the number of years since birth. While one typically refers to a newborn as being one year of age, for the purpose of mortality, these children are reported to be under 1 year of age, and so forth. The ages for which data were available were as follows: under 1 year, 1 year, 2 years, 3 years, 4 years, under 5 years, 5-9 years, 10-14 years, 15-19 years, 20-24 years, 25-29 years, 30-34 years, 35-39 years, 40-44 years, 45-49 years, 50-54 years, 55-59 years, 60-64 years, 65-69 years, 70-74 years, 75-79 years, 80-84 years, 85-89 years, 90-94 years, 95-99 years, 100+ years, and unknown age. For the reporting years of 1900-1909 the category for 100+ years was not available, but 95+ was used instead.[1] Although not used herein, mortality rates were reported for the years 1880 and 1890 in their respective Census reports, but no intercensus data is available.

Mistakes are inevitable when transcribing so many numbers from one source to another. For quality assurance, the values listed in the TOTAL columns were not actually transcribed. Instead, the Microsoft Excel™ spreadsheets were directed to sum up each of the rows, representing the independent gender and race cohorts. This only assures that the total of the numbers transcribed were correct. To assure that the correct number was placed in the correct cell, column totals were also summed by the spreadsheet. In doing so, one can catch transcribing errors if numbers were accidentally shifted within a row. (This was a predominant problem for diseases with many blank cells). Figure 13 demonstrates how to calculate these totals (Cell references: letter represents column, and number represents the row of the cell).

---

[1] For the reporting years of 1900-1913, race was not delineated. In 1914, 95.5% of the population of the states and counties with death registries was reported to be of White descent, so the 1900-1913 data is used herein as an estimate for European American mortality. For the years 1914-1932, no distinction was made between Black Americans and Other Americans.

In 1972, the final mortality numbers were not released. The U.S. Department of Health and Human Services opted to report numbers for half of the death records and multiplying the results by two. In 1962 and 1963, mortality data was published without values for New Jersey, which were added only after publication. Corrected mortality for these years were available only for the 5-year age groups of 0-4 years to 80-84-years (U.S. Department of Health and Human Services, 1982b).

The diseases for which data were collected are as follows (ICD-9 Number):

All causes

- I. Infectious and parasitic diseases (001-139)
  - Intestinal infectious diseases (001-009)
    - Intestinal infections due to other specified organisms (007-008)
    - Ill-defined infections (009)
  - Viral hepatitis (070)
- II. Neoplasms (140-239)
  - Malignant neoplasms, including neoplasms of lymphatic and hematopoietic tissues (140-208)
    - Malignant neoplasms of lip, oral cavity and pharynx (140-149)
      - Of lip (140)
      - Of tongue (141)
      - Of pharynx (146-149.0)
      - Of other and ill-defined sites within the lip, oral cavity, and pharynx (142-145, 149.1-149.9)
    - Malignant neoplasms of digestive organs and peritoneum (150-159)
      - Of esophagus (150)
      - Of stomach (151)
      - Of small intestine, including duodenum (152)
      - Of colon (153)
        - Hepatic and splenic flexures and transverse colon (153.0-153.1, 153.7)
        - Descending colon (153.2)
        - Sigmoid colon (153.3)
        - Cecum, appendix, and ascending colon (153.4-153.6)
        - Other and colon, unspecified (153.8-153.9)
      - Of rectum, rectosigmoid junction, and anus (154)
      - Of liver and intrahepatic bile ducts (155)
        - Liver, primary (155.0)
        - Intrahepatic bile ducts (155.1)
        - Liver, not specified as primary or secondary (155.2)
      - Of gallbladder and extrahepatic bile ducts (156)
      - Of pancreas (157)
      - Of retroperitoneum, peritoneum, and other and ill-defined sites within the digestive organs and peritoneum (158-159)
    - Malignant neoplasms of respiratory and intrathoracic organs (160-165)
      - Of larynx (161)
      - Of trachea, bronchus, and lung (162)
      - Of all other and ill-defined sites within the respiratory system and intrathoracic organs (160, 163-165)
    - Malignant neoplasms of bone, connective tissue, skin, and breast (170-175)
      - Of bone and articular cartilage (170)

Of connective and other soft tissue (171)
Melanoma of skin (172)
Other malignant neoplasms of skin (173)
Of female breast (174)
Of male breast (175)
Malignant neoplasms of genital organs (179-187)
Of cervix uteri (180)
Of other parts of uterus (179, 181-182)
Of ovary and other uterine adnexa (183)
Of other and unspecified female genital organs (184)
Of prostate (185)
Of testis (186)
Of penis and other male genital organs (187)
Malignant neoplasms of urinary organs (188-189)
Of bladder (188)
Of kidney and other and unspecified urinary organs (189)
Malignant neoplasms of other and unspecified sites (190-199)
Of eye (190)
Of brain (191)
Of other and unspecified parts of nervous system (192)
Of thyroid gland and other endocrine glands and related structures (193-194)
Of all other and unspecified sites (195-199)
Malignant neoplasms of lymphatic and hematopoietic tissues (200-208)
Lymphosarcoma and reticulosarcoma (200)
Hodgkin's disease (201)
Other malignant neoplasms of lymphoid and histiocytic tissue (202)
Multiple myeloma and immunoproliferative neoplasms (203)
Leukemia (204-208)
Lymphoid leukemia (204)
Myeloid leukemia (205)
Monocytic leukemia (206)
Other and unspecified leukemia (207-208)
Benign neoplasms, carcinoma in situ, and neoplasms of uncertain behavior and of unspecified nature (210-239)
Benign neoplasms (210-229)
Of female genital organs (218-221)
Of eye, brain and other parts of nervous system (224-225)
Of all other and unspecified sites (210-217, 222-223, 226-229)
Carcinoma in situ (230-234)
Of breast and genitourinary system (233)
Of all other and unspecified sites (230-232, 234)
Neoplasms of uncertain behavior (235-238)
Neoplasms of unspecified nature (239)
III. Endocrine, nutritional, and metabolic diseases and immunity disorders (240-279)

- Disorders of thyroid gland (240-246)
- Diabetes mellitus (250)
- Cystic fibrosis (277.0)

IV. Diseases of blood and blood-forming organs (280-289)
V. Mental disorders (290-319)
- Senile and presenile organic psychotic conditions (290)

VI. Diseases of the nervous system and sense organs (320-389)
- Multiple sclerosis (340)

VII. Diseases of the circulatory system (390-459)
- Hypertensive disease (401-404)
- Ischemic heart disease (410-414)
- Cerebrovascular diseases (430-438)
- Atherosclerosis (440)

VIII. Diseases of the respiratory system (460-519)
- Acute bronchitis and bronchiolitis (466)
- Influenza (487)
- Chronic obstructive pulmonary diseases and allied conditions (490-496)
  - Bronchitis, chronic and unspecified, emphysema, and asthma (490-493)
    - Bronchitis, not specified as acute or chronic (490)
    - Chronic bronchitis (491)
    - Emphysema (492)
    - Asthma (493)
    - Bronchiectasis and extrinsic allergic alveolitis (494-495)
  - Chronic airways obstruction, not elsewhere classified (496)

IX. Diseases of the digestive system (520-579)
- Ulcer of stomach and duodenum (531-533)
- Chronic liver disease and cirrhosis (571)

X. Diseases of the genitourinary system (580-629)
XII. Diseases of the skin and subcutaneous tissue (680-709)
XIII. Diseases of the musculoskeletal system and connective tissue (710-739)
XIV. Congenital anomalies (740-759)
XV. Certain conditions originating in the perinatal period (760-779)
XVI. Symptoms, signs, and ill-defined conditions (780-799)
- Senility without mention of psychosis (797)

Accidents and adverse effects (E800-E949)
Suicide (E950-E959)
Homicide and legal intervention (E960-E978)
Injury undetermined whether accidentally or purposely inflicted (E980-E989)
Injury resulting from operations of war (E990-E999)

Files for each disease can be retrieved at our website (See end of ABSTRACT for instructions.)

3.2.2 Population Data

Population values were provided by the Duke Center for Demographic Studies for the years 1950 to 1992. For previous years, population estimates were derived directly from census counts for those states and counties reporting to the national death registries. The states with death registries are reported at the beginning of the mortality data books (U.S. Bureau of the Census, 1900-1933; Texas was the last existing state to be added to U.S. mortality results in 1933; Hawaii and Alaska were added later when they became states).

Population values are available for the same age groups as in the mortality data sets for reporting years 1950-1992. Prior to 1950, population values for children are only available for the age groups: under 1 year, and 1-4 year olds.

3.2.3 Mortality Rate Data

Combining mortality and population data sets permit calculation of the age-specific mortality rate for each birth year, "h", designated OBS(h,t) for the "OBServed" form of death mortality rate at age "t".

(Eq. 1)

$$OBS(h,t) = \frac{\text{recorded deaths from the form of death of interest for birth cohort h at age t}}{\text{recorded population size from birth cohort h at age t}}$$

An individual who is listed as being X years old will turn (X + 1) years old during the year. Yearly mortality data lists individuals by their age at the time of death, such that this individual may be listed as dying at age (X + 1) in the mortality data set, but is listed as being X

years old in the population data set, since the population estimates were for the beginning of the year.

To correct for this confounding factor, the mortality values of, for example, reporting year 1950, are linked with the population values of the year 1951. These population values would represent the number of individuals who survived during the year 1950, to which is added the number of deaths during the year 1950. For example, OBS(1900,50) would be derived by determining the number of people who died in 1950 at age of 50 from the form of death of interest, and dividing that by the number of 51 year olds alive at the beginning of 1951 plus the number of **all** 50-year olds who died in 1950. The 51-year olds alive in 1951 and the number of 50-year olds who died in 1950, represent the individuals who could have potentially or actually died at age 50 during 1950.

With this correction, any individual who was 50 years of age at the beginning of the year, turns 51 and dies, would be included in the number of deaths occurring at age 51, and he/she is included under the population of 51 year-olds, thereby guaranteeing that both numerator and denominator of the calculated OBServed mortality rates consist of the same pool of individuals.

To convert our extensive database of raw mortality numbers into mortality rates required the construction of a template spreadsheet to more rapidly do all calculations. The template that has been built has been extended to the reporting year of 2050. This template was used by Jose Marquez, MS, when putting together the mortality data for Japan for the reporting years of 1950-1995. (See end of ABSTRACT for instructions on acquiring this and any other file of interest). As new mortality and population values are collected, they need only be added to the template file which will automatically convert these values into rates, grouped by birth-year and decade cohorts. Furthermore, all relevant graphs are generated by this template.

Figures 14 through 20 demonstrate how to construct such a template (Some knowledge of Microsoft Excel™ or the spreadsheet program of choice is needed). A template file is made up of multiple worksheets presenting data and making calculations in the desired methods. To travel from one worksheet to another, one need only click on the name of the worksheet as listed on the lower left corner of the template file. For example, Figure 14 shows that the Raw Data (EAM), Raw Data (EAF), Raw Data (NEAM), and Raw Data (NEAF) worksheets are available. To look at any other existing worksheets not listed one would click on the arrows on the lower left corner of the active window.

The first four worksheets of the template file (Figure 14) consist of the raw numbers of deaths as already typed into the raw data spreadsheet shown in Figure 13, using the sum of the numbers for Black and Other Americans to calculate the number of deaths for non-European Americans. (One could conceivably construct 6 separate worksheets to represent each gender and race combination).

Each row of these worksheets consists of the numbers of deaths for each reporting year for the form of death of interest; each column represents the number of deaths for each age group (as listed in the original raw data file). Scrolling down in the active window will reveal the rows for the reporting year of interest for which data needs to be transcribed as data becomes available for more years.

It is advisable to place any cautionary notes in these files. For example, Figure 14 reveals that for the reporting years of 1900-1909, leukemia was not included as a cancerous form of death. During these years, leukemia had been listed under the form of death, "Anemias and leukemias". Mortality data for the next year, 1910, revealed that more than 90% of the sum of

deaths by "Anemias and leukemias", were in fact anemias. For larger accuracy, leukemia values were excluded prior to 1910.

The next four worksheets (Figure 15) provide the equivalent population values for each reporting mortality year. Population values in these worksheets have already been adjusted as mentioned at the beginning of Section 3.2.3. In the Population (NEAM) and Population (NEAF) worksheets there is a note to remind the user that populations for Non-European Americans were combined with European Americans for the years 1900-1913, as mortality numbers were not delineated by race for these years.

The next four worksheets (Figures 16, 17) calculate the mortality rate for each birthyear cohort and each age group. Ages are listed by the midpoint of the age groups (i.e. 9-14 years becomes 12.5). Figure 16 provides the calculated rates, and Figure 17 provides the equations used to do so.[2] For example, the calculated cancer mortality rate of 0.5 year old (under 1 year) European American males born in 1900 is equal (=) to the number of deaths in this age group during the year 1901 (cell reference: 'Raw Data (EAM)'!C4 as shown in Figure 14) divided (/) by the equivalent population in 1901 (cell reference: 'Population (EAM)'!C4 as shown in Figure 15). Representation of mortality/incidence rates per 100,000 requires further multiplication (* 10^5).

The next two worksheets (Figures 18, 19) recalculate the mortality rates, but now as birth decade cohorts. 1800s represents individuals born in 1800-1809 and so forth. The formulas used to calculate these rates are similar to those for individual birthyears (as shown in Figure 17), with the addition of the SUM function for use with 10 reporting years' worth of data. The number of

[2] NOTE: As can be observed, at the time that this template was generated, we made the decision to transpose the data such that the mortality rate for each age group is now listed by row instead of by column (compare Figure 15 with Figure 16). Anyone familiar with the formula filling features of Microsoft Excel™ knows that this template could have been generated far more easily without having done so.

cancer deaths among 0.5 year olds born in the 1900s would have occurred in the reporting years 1901-1910 (cell references: SUM('Raw Data (EAM)'!C4:C13) as shown in Figure 14).

Note that it is more appropriate to calculate birthyear decade mortality rates by dividing the sum of the number of deaths during 10 years by the sum of the populations during these 10 years. It is not accurate to alternately take a 10-year average of single birthyear mortality rates.

The following worksheets (Figures 20-22) contain pregenerated graphs for the mortality rates calculated in the previous worksheets. These graphs are automatically updated when new data is introduced to the raw data worksheets. Figure 20 illustrates mortality curves by birth decade cohort for all age groups. Figure 21 shows mortality curves, but only up to age 35, to facilitate the observation of childhood cancer rates. The template file also contains pregenerated graphs that break down these mortality curves into the age ranges of 35-80 and 80 to above, not shown.

Figure 22 shows an alternate way by which to plot mortality rates for each individual age group as a function of the birthyear cohort. This is relatively useful when assessing any potential unknown changes in disease classification or treatment. The template contains such plots for the following age groups (0.5, 3, 12.5, 22.5, 32.5, 42.5, 52.5, 62.5, 72.5, 82.5, 92.5, and 102.5).

For example, mortality by diabetes mellitus among 52.5 year old European Americans suddenly dropped for the birthyear cohort of 1877 (Figure 23). The exact time point of this drop corresponds to the year 1877 + 52 = 1949. This drop is actually not due to the advent of insulin, but is actually a change in the classification of death by diabetes. Prior to 1949, deaths by heart disease among the diabetic were reported as being due to diabetes. Since 1949, deaths by heart disease among the diabetic are reported as being due to heart disease instead.

Fig. 14: MORTALITY TEMPLATE: Raw mortality data

Number of deaths by all forms of neoplasms, reported by year of death and age at death. Age at death corresponds to the years since birth.

Microsoft Excel - Neoplasms.1

File Edit View Insert Format Tools Data Window Help

Mortality by all Neoplasms	Total	under one year	1 year	2 years	3 years	4 years	under 5 years	5-9 years
excludes leukemia 1900-1909								
1900	7,451	20	15	14	12	8	69	31
1901	7,857	20	10	16	13	14	73	29
1902	7,921	21	8	16	15	8	68	31
1903	8,574	18	13	13	15	10	69	33
1904	8,994	26	9	9	17	11	72	29
1905	9,326	23	9	12	10	13	67	33
1906	11,281	27	11	14	20	13	85	43
1907	11,940	21	19	22	15	7	84	29
1908	13,163	20	14	23	16	11	84	31
1909	15,046	31	12	20	20	18	101	45
1910	17,057	45	26	49	24	30	174	73
1911	18,189	47	37	28	33	20	165	73
1912	19,175	49	39	30	46	36	200	106
1913	20,771	38	38	44	43	31	194	103
1914	21,515	42	32	50	41	28	193	108
1915	22,716	56	45	56	36	42	235	123
1916	24,185	50	55	35	48	39	227	117
1917	25,470	66	50	42	40	35	233	140
1918	26,705	62	62	55	41	38	258	108
1919	28,551	46	52	44	58	42	242	129
1920	30,457	59	67	44	49	44	263	141
1921	32,246	72	49	61	63	51	296	139
1922	34,780	55	72	66	53	55	301	140
1923	37,546	76	72	68	63	48	327	178
1924	39,410	67	64	79	65	60	335	226
1925	41,267	57	75	82	85	57	356	206
1926	43,590	60	81	87	81	54	363	220
1927	44,911	75	66	68	89	66	364	226
1928	48,364	77	74	90	83	66	390	255
1929	49,395	64	61	93	81	78	377	248
1930	52,590	76	70	94	109	100	449	322
1931	53,797	72	72	107	103	95	449	314
1932	56,107	78	68	103	92	89	430	335
1933	59,939	76	78	93	106	106	459	336
1934	62,807	94	100	85	112	108	499	364
1935	64,515	74	86	109	105	87	461	331
1936	67,223	90	88	84	110	92	464	382
1937	68,993	95	93	124	91	122	525	383
1938	71,506	88	114	121	118	106	547	358
1939	72,603	107	108	122	101	84	522	329
1940	75,674	126	100	130	132	114	602	372
1941	76,583	122	120	156	141	116	655	371
1942	78,026	86	104	132	146	99	567	340
1943	79,744	106	120	148	148	114	636	371
1944	82,270	105	147	156	149	96	653	401
1945	85,725	106	127	168	161	134	696	401
1946	88,518	126	128	158	194	139	745	405
1947	93,276	126	130	182	178	176	792	457
1948	96,962	125	171	169	189	141	795	463

Raw Data (EAM) Raw Data (EAF) Raw Data (NEAM) Raw Data (NEAF)

Fig. 15: MORTALITY TEMPLATE: Raw population data

Population values are estimated as in the following example.

Number of (5-9) years olds during the year 1900 = Number of (5-9) year olds alive at the beginning of 1901 plus the number of (5-9) year olds who died during 1900.

Microsoft Excel - Neoplasms.1

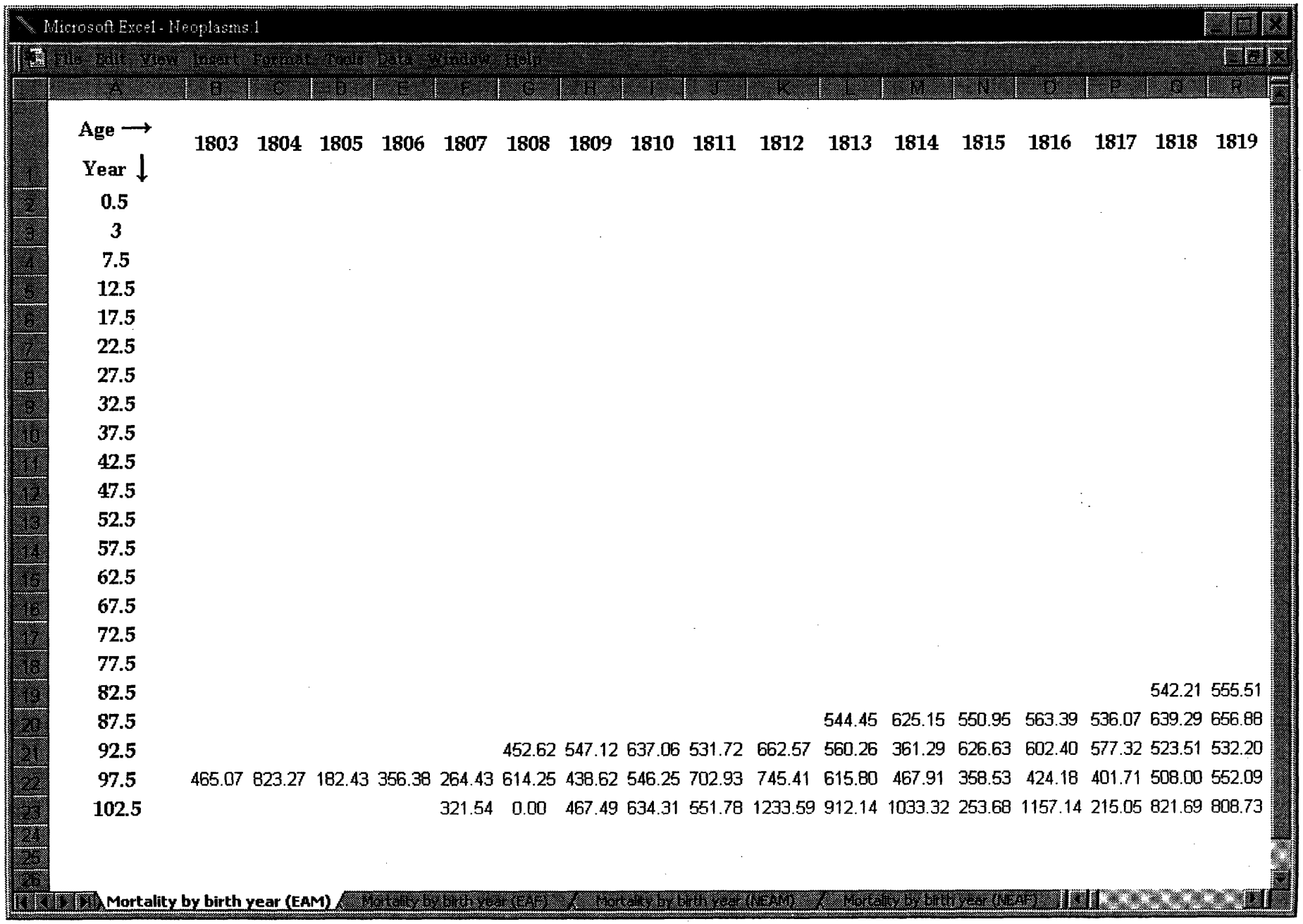
File Edit View Insert Format Tools Data Window Help

Age → / Year ↓	All ages	under one year	1-4 years	5-9 years	10-14 years	15-19 years	20-24 years
1900	10,460,670	287,058	864,149	1,016,916	931,529	910,184	951,874
1901	10,665,737	282,744	873,499	1,026,670	946,249	931,251	976,991
1902	10,878,683	286,992	887,406	1,037,319	960,987	952,435	1,002,038
1903	11,101,311	289,733	899,797	1,047,972	976,354	974,149	1,027,616
1904	11,331,931	296,890	913,632	1,058,326	991,520	995,695	1,053,043
1905	11,544,247	302,024	926,504	1,068,475	1,006,181	1,016,791	1,078,124
1906	18,003,611	459,698	1,459,915	1,673,780	1,572,832	1,597,082	1,706,471
1907	18,390,526	465,091	1,483,490	1,692,439	1,596,436	1,631,674	1,751,100
1908	20,569,525	511,643	1,653,526	1,885,327	1,788,456	1,835,225	1,972,997
1909	23,441,331	570,773	1,875,114	2,127,683	2,029,548	2,099,450	2,253,682
1910	25,257,297	612,104	2,032,574	2,316,035	2,199,245	2,235,926	2,397,635
1911	28,552,932	680,630	2,333,125	2,679,178	2,535,302	2,527,375	2,657,156
1912	28,980,075	683,660	2,371,144	2,732,320	2,577,040	2,532,031	2,653,275
1913	30,522,033	722,358	2,523,915	2,922,057	2,746,011	2,650,155	2,749,560
1914	30,407,088	708,339	2,531,764	2,938,011	2,753,247	2,616,888	2,683,965
1915	30,824,471	708,772	2,568,193	2,990,802	2,795,934	2,621,253	2,677,832
1916	32,387,568	751,595	2,740,991	3,196,401	2,976,141	2,740,348	2,766,649
1917	33,778,300	781,908	2,876,930	3,370,799	3,133,014	2,840,241	2,838,869
1918	38,510,413	876,431	3,277,669	3,853,010	3,574,694	3,207,291	3,193,606
1919	39,668,929	882,985	3,383,178	4,014,675	3,722,422	3,283,407	3,240,993
1920	41,056,604	900,193	3,444,606	4,143,446	3,858,337	3,432,117	3,363,783
1921	41,726,529	891,656	3,439,601	4,201,012	3,927,333	3,520,054	3,427,614
1922	43,689,132	918,611	3,555,738	4,404,056	4,134,500	3,729,818	3,599,427
1923	45,682,432	942,020	3,655,681	4,583,457	4,320,789	3,931,657	3,770,520
1924	46,742,009	944,833	3,684,900	4,681,407	4,429,046	4,057,976	3,867,911
1925	48,189,643	961,916	3,755,489	4,831,932	4,578,456	4,219,664	3,993,125
1926	49,113,238	963,507	3,768,758	4,906,733	4,663,583	4,327,546	4,076,399
1927	50,531,264	970,783	3,826,665	5,053,207	4,816,412	4,497,054	4,205,383
1928	53,338,013	1,017,317	4,004,606	5,351,101	5,109,435	4,804,729	4,461,822
1929	54,331,625	1,016,389	4,019,419	5,435,995	5,202,772	4,924,914	4,554,467
1930	54,612,075	1,004,826	3,976,044	5,341,675	5,181,609	4,957,311	4,587,550
1931	54,904,077	989,703	3,933,968	5,247,745	5,160,981	4,990,419	4,620,327
1932	55,196,733	975,360	3,891,387	5,153,275	5,140,168	5,022,606	4,652,830
1933	57,941,007	1,010,709	4,036,474	5,304,291	5,360,516	5,296,955	4,907,445
1934	58,326,812	1,006,574	3,999,671	5,212,729	5,345,115	5,335,696	4,944,722
1935	58,691,560	996,520	3,960,380	5,120,795	5,329,821	5,374,288	4,981,643
1936	59,089,274	988,752	3,922,815	5,028,650	5,313,964	5,413,137	5,018,674
1937	59,436,861	979,024	3,884,276	4,935,999	5,297,687	5,451,101	5,054,650
1938	59,756,781	969,298	3,845,635	4,843,149	5,281,014	5,487,781	5,089,172

Population (EAM) | Population (EAF) | Population (NEAM) | Population (NEAF)

Fig. 16: MORTALITY TEMPLATE: Calculated age-specific cancer mortality rates by year of birth.

(Reported per 100,000)

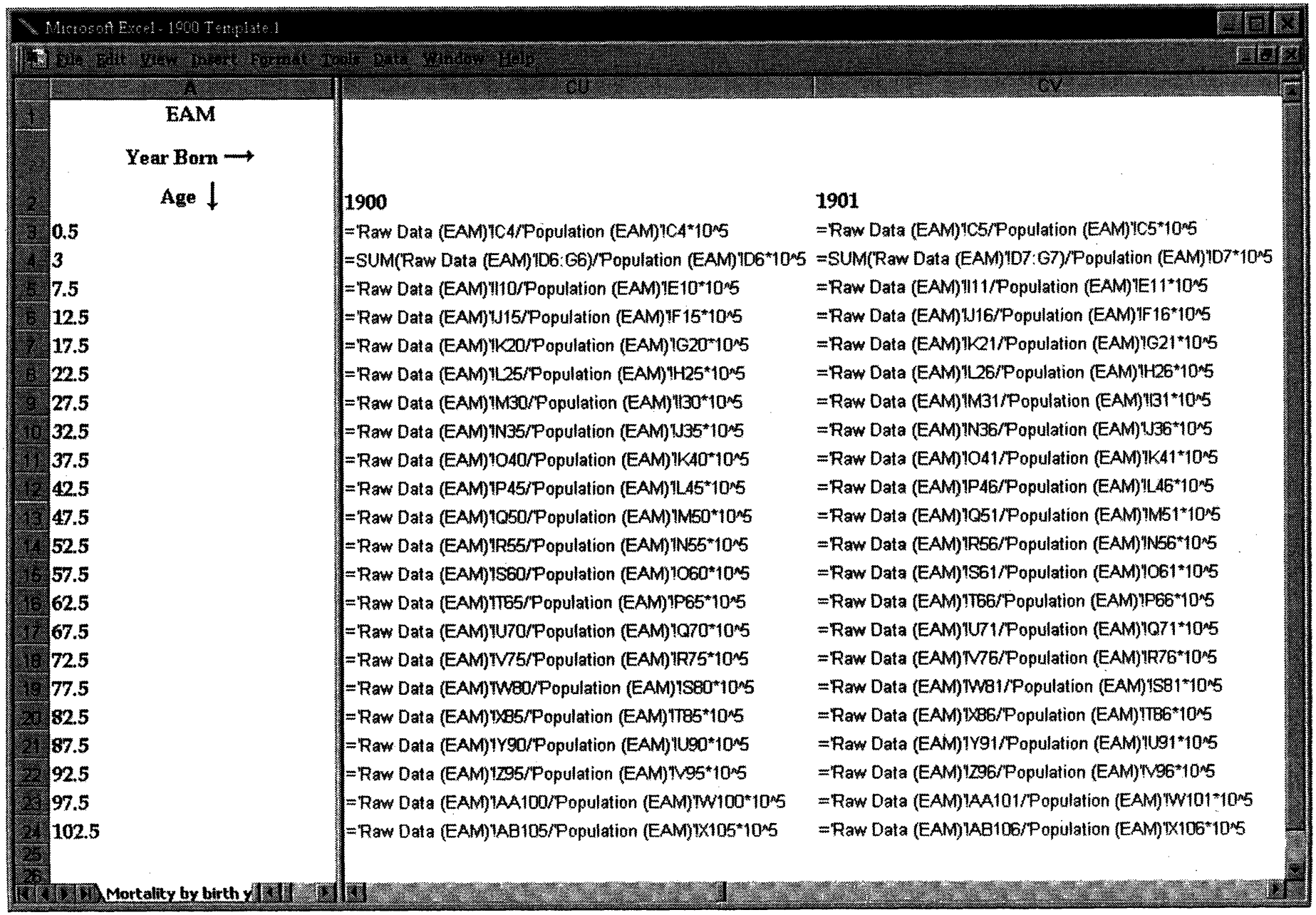


Age → / Year ↓	1803	1804	1805	1806	1807	1808	1809	1810	1811	1812	1813	1814	1815	1816	1817	1818	1819
0.5																	
3																	
7.5																	
12.5																	
17.5																	
22.5																	
27.5																	
32.5																	
37.5																	
42.5																	
47.5																	
52.5																	
57.5																	
62.5																	
67.5																	
72.5																	
77.5																	
82.5																542.21	555.51
87.5											544.45	625.15	550.95	563.39	536.07	639.29	656.88
92.5						452.62	547.12	637.06	531.72	662.57	560.26	361.29	626.63	602.40	577.32	523.51	532.20
97.5	465.07	823.27	182.43	356.38	264.43	614.25	438.62	546.25	702.93	745.41	615.80	467.91	358.53	424.18	401.71	508.00	552.09
102.5					321.54	0.00	467.49	634.31	551.78	1233.59	912.14	1033.32	253.68	1157.14	215.05	821.69	808.73

Fig. 17: MORTALITY TEMPLATE: Calculated age-specific cancer mortality rates by year of birth w/ formulae

(Reported per 100,000)

Example: Calculated cancer mortality rate of 0.5 year old (under 1 year) European American males born in 1900 is equal (=) to number of deaths in this age group during the year 1901 (cell reference: 'Raw Data (EAM)'!C4 as shown in Figure 14) divided (/) by the equivalent population in 1901 (cell reference: 'Population (EAM)'!C4 as shown in Figure 15) reported per 100,000 (*10^5).



Microsoft Excel - 1900 Template 1

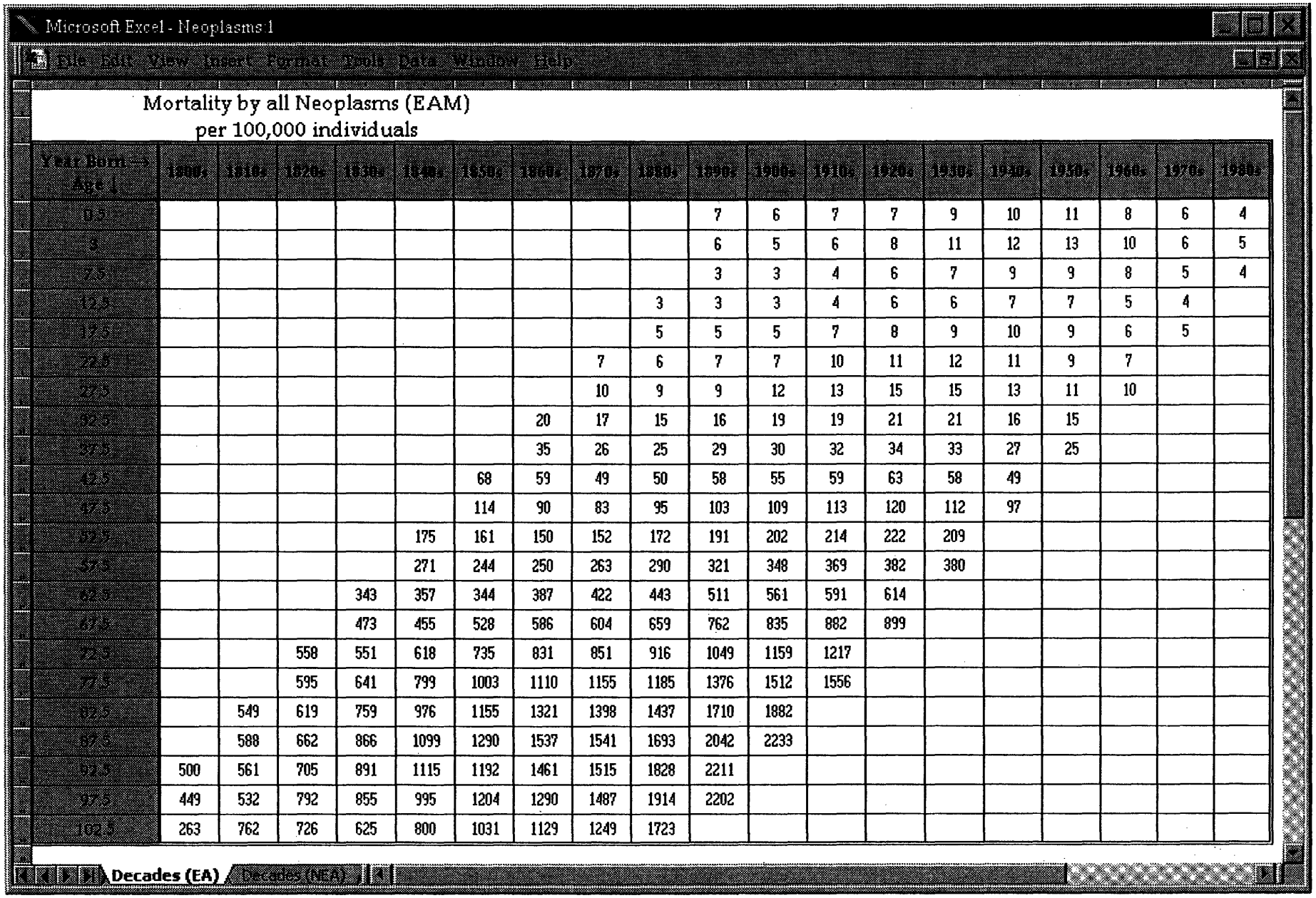
File Edit View Insert Format Tools Data Window Help

	A	CU	CV
1	EAM		
	Year Born →		
2	Age ↓	1900	1901
3	0.5	='Raw Data (EAM)'!C4/'Population (EAM)'!C4*10^5	='Raw Data (EAM)'!C5/'Population (EAM)'!C5*10^5
4	3	=SUM('Raw Data (EAM)'!D6:G6)/'Population (EAM)'!D6*10^5	=SUM('Raw Data (EAM)'!D7:G7)/'Population (EAM)'!D7*10^5
5	7.5	='Raw Data (EAM)'!I10/'Population (EAM)'!E10*10^5	='Raw Data (EAM)'!I11/'Population (EAM)'!E11*10^5
6	12.5	='Raw Data (EAM)'!J15/'Population (EAM)'!F15*10^5	='Raw Data (EAM)'!J16/'Population (EAM)'!F16*10^5
7	17.5	='Raw Data (EAM)'!K20/'Population (EAM)'!G20*10^5	='Raw Data (EAM)'!K21/'Population (EAM)'!G21*10^5
8	22.5	='Raw Data (EAM)'!L25/'Population (EAM)'!H25*10^5	='Raw Data (EAM)'!L26/'Population (EAM)'!H26*10^5
9	27.5	='Raw Data (EAM)'!M30/'Population (EAM)'!I30*10^5	='Raw Data (EAM)'!M31/'Population (EAM)'!I31*10^5
10	32.5	='Raw Data (EAM)'!N35/'Population (EAM)'!J35*10^5	='Raw Data (EAM)'!N36/'Population (EAM)'!J36*10^5
11	37.5	='Raw Data (EAM)'!O40/'Population (EAM)'!K40*10^5	='Raw Data (EAM)'!O41/'Population (EAM)'!K41*10^5
12	42.5	='Raw Data (EAM)'!P45/'Population (EAM)'!L45*10^5	='Raw Data (EAM)'!P46/'Population (EAM)'!L46*10^5
13	47.5	='Raw Data (EAM)'!Q50/'Population (EAM)'!M50*10^5	='Raw Data (EAM)'!Q51/'Population (EAM)'!M51*10^5
14	52.5	='Raw Data (EAM)'!R55/'Population (EAM)'!N55*10^5	='Raw Data (EAM)'!R56/'Population (EAM)'!N56*10^5
15	57.5	='Raw Data (EAM)'!S60/'Population (EAM)'!O60*10^5	='Raw Data (EAM)'!S61/'Population (EAM)'!O61*10^5
16	62.5	='Raw Data (EAM)'!T65/'Population (EAM)'!P65*10^5	='Raw Data (EAM)'!T66/'Population (EAM)'!P66*10^5
17	67.5	='Raw Data (EAM)'!U70/'Population (EAM)'!Q70*10^5	='Raw Data (EAM)'!U71/'Population (EAM)'!Q71*10^5
18	72.5	='Raw Data (EAM)'!V75/'Population (EAM)'!R75*10^5	='Raw Data (EAM)'!V76/'Population (EAM)'!R76*10^5
19	77.5	='Raw Data (EAM)'!W80/'Population (EAM)'!S80*10^5	='Raw Data (EAM)'!W81/'Population (EAM)'!S81*10^5
20	82.5	='Raw Data (EAM)'!X85/'Population (EAM)'!T85*10^5	='Raw Data (EAM)'!X86/'Population (EAM)'!T86*10^5
21	87.5	='Raw Data (EAM)'!Y90/'Population (EAM)'!U90*10^5	='Raw Data (EAM)'!Y91/'Population (EAM)'!U91*10^5
22	92.5	='Raw Data (EAM)'!Z95/'Population (EAM)'!V95*10^5	='Raw Data (EAM)'!Z96/'Population (EAM)'!V96*10^5
23	97.5	='Raw Data (EAM)'!AA100/'Population (EAM)'!W100*10^5	='Raw Data (EAM)'!AA101/'Population (EAM)'!W101*10^5
24	102.5	='Raw Data (EAM)'!AB105/'Population (EAM)'!X105*10^5	='Raw Data (EAM)'!AB106/'Population (EAM)'!X106*10^5
25			
26			

Mortality by birth y

Fig. 18: MORTALITY TEMPLATE: Calculated age-specific cancer mortality rates by decade of birth.

(Reported per 100,000)



Microsoft Excel - Neoplasms 1

File Edit View Insert Format Tools Data Window Help

Mortality by all Neoplasms (EAM)
per 100,000 individuals

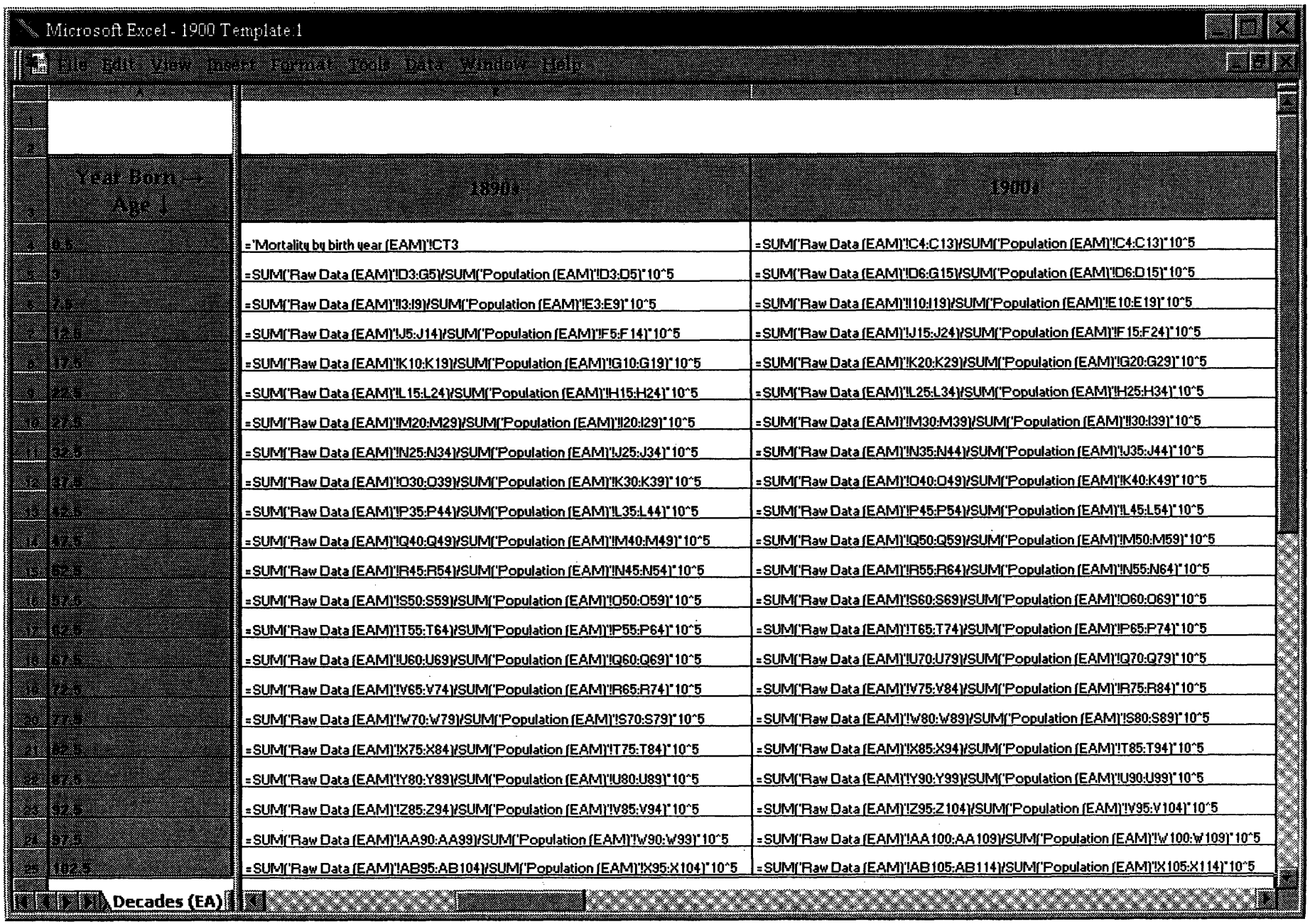
Year Born → / Age ↓	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s	1890s	1900s	1910s	1920s	1930s	1940s	1950s	1960s	1970s	1980s
0.5										7	6	7	7	9	10	11	8	6	4
3										6	5	6	8	11	12	13	10	6	5
7.5										3	3	4	6	7	9	9	8	5	4
12.5									3	3	3	4	6	6	7	7	5	4	
17.5									5	5	5	7	8	9	10	9	6	5	
22.5								7	6	7	7	10	11	12	11	9	7		
27.5								10	9	9	12	13	15	15	13	11	10		
32.5							20	17	15	16	19	19	21	21	16	15			
37.5							35	26	25	29	30	32	34	33	27	25			
42.5						68	59	49	50	58	55	59	63	58	49				
47.5						114	90	83	95	103	109	113	120	112	97				
52.5					175	161	150	152	172	191	202	214	222	209					
57.5					271	244	250	263	290	321	348	369	382	380					
62.5				343	357	344	387	422	443	511	561	591	614						
67.5				473	455	528	586	604	659	762	835	882	899						
72.5			558	551	618	735	831	851	916	1049	1159	1217							
77.5			595	641	799	1003	1110	1155	1185	1376	1512	1556							
82.5		549	619	759	976	1155	1321	1398	1437	1710	1882								
87.5		588	662	866	1099	1290	1537	1541	1693	2042	2233								
92.5	500	561	705	891	1115	1192	1461	1515	1828	2211									
97.5	449	532	792	855	995	1204	1290	1487	1914	2202									
102.5	263	762	726	625	800	1031	1129	1249	1723										

Decades (EA) / Decades (NEA)

Fig. 19: MORTALITY TEMPLATE: Calculated age-specific cancer mortality rates by decade of birth w/ formulae

(Reported per 100,000)

Example: Calculated cancer mortality rate of 0.5 year old (under 1 year) European American males born in the 1900s is equal (=) to the sum of all deaths among 0.5 year olds born in the occurring in the reporting years 1901-1910 (cell references: SUM('Raw Data (EAM)'!C4:C13) as shown in Figure 14) divided (/) by the equivalent population in 1901-1910 (cell references: SUM('Population (EAM)'!C4:C13) as shown in Figure 15) reported per 100,000 (*10^5).



Microsoft Excel - 1900 Template.1

File Edit View Insert Format Tools Data Window Help

Year Born → Age ↓	1890s	1900s
0.5	='Mortality by birth year (EAM)'!CT3	=SUM('Raw Data (EAM)'!C4:C13)/SUM('Population (EAM)'!C4:C13)*10^5
3	=SUM('Raw Data (EAM)'!D3:G5)/SUM('Population (EAM)'!D3:D5)*10^5	=SUM('Raw Data (EAM)'!D6:G15)/SUM('Population (EAM)'!D6:D15)*10^5
7.5	=SUM('Raw Data (EAM)'!I3:I9)/SUM('Population (EAM)'!E3:E9)*10^5	=SUM('Raw Data (EAM)'!I10:I19)/SUM('Population (EAM)'!E10:E19)*10^5
12.5	=SUM('Raw Data (EAM)'!J5:J14)/SUM('Population (EAM)'!F5:F14)*10^5	=SUM('Raw Data (EAM)'!J15:J24)/SUM('Population (EAM)'!F15:F24)*10^5
17.5	=SUM('Raw Data (EAM)'!K10:K19)/SUM('Population (EAM)'!G10:G19)*10^5	=SUM('Raw Data (EAM)'!K20:K29)/SUM('Population (EAM)'!G20:G29)*10^5
22.5	=SUM('Raw Data (EAM)'!L15:L24)/SUM('Population (EAM)'!H15:H24)*10^5	=SUM('Raw Data (EAM)'!L25:L34)/SUM('Population (EAM)'!H25:H34)*10^5
27.5	=SUM('Raw Data (EAM)'!M20:M29)/SUM('Population (EAM)'!I20:I29)*10^5	=SUM('Raw Data (EAM)'!M30:M39)/SUM('Population (EAM)'!I30:I39)*10^5
32.5	=SUM('Raw Data (EAM)'!N25:N34)/SUM('Population (EAM)'!J25:J34)*10^5	=SUM('Raw Data (EAM)'!N35:N44)/SUM('Population (EAM)'!J35:J44)*10^5
37.5	=SUM('Raw Data (EAM)'!O30:O39)/SUM('Population (EAM)'!K30:K39)*10^5	=SUM('Raw Data (EAM)'!O40:O49)/SUM('Population (EAM)'!K40:K49)*10^5
42.5	=SUM('Raw Data (EAM)'!P35:P44)/SUM('Population (EAM)'!L35:L44)*10^5	=SUM('Raw Data (EAM)'!P45:P54)/SUM('Population (EAM)'!L45:L54)*10^5
47.5	=SUM('Raw Data (EAM)'!Q40:Q49)/SUM('Population (EAM)'!M40:M49)*10^5	=SUM('Raw Data (EAM)'!Q50:Q59)/SUM('Population (EAM)'!M50:M59)*10^5
52.5	=SUM('Raw Data (EAM)'!R45:R54)/SUM('Population (EAM)'!N45:N54)*10^5	=SUM('Raw Data (EAM)'!R55:R64)/SUM('Population (EAM)'!N55:N64)*10^5
57.5	=SUM('Raw Data (EAM)'!S50:S59)/SUM('Population (EAM)'!O50:O59)*10^5	=SUM('Raw Data (EAM)'!S60:S69)/SUM('Population (EAM)'!O60:O69)*10^5
62.5	=SUM('Raw Data (EAM)'!T55:T64)/SUM('Population (EAM)'!P55:P64)*10^5	=SUM('Raw Data (EAM)'!T65:T74)/SUM('Population (EAM)'!P65:P74)*10^5
67.5	=SUM('Raw Data (EAM)'!U60:U69)/SUM('Population (EAM)'!Q60:Q69)*10^5	=SUM('Raw Data (EAM)'!U70:U79)/SUM('Population (EAM)'!Q70:Q79)*10^5
72.5	=SUM('Raw Data (EAM)'!V65:V74)/SUM('Population (EAM)'!R65:R74)*10^5	=SUM('Raw Data (EAM)'!V75:V84)/SUM('Population (EAM)'!R75:R84)*10^5
77.5	=SUM('Raw Data (EAM)'!W70:W79)/SUM('Population (EAM)'!S70:S79)*10^5	=SUM('Raw Data (EAM)'!W80:W89)/SUM('Population (EAM)'!S80:S89)*10^5
82.5	=SUM('Raw Data (EAM)'!X75:X84)/SUM('Population (EAM)'!T75:T84)*10^5	=SUM('Raw Data (EAM)'!X85:X94)/SUM('Population (EAM)'!T85:T94)*10^5
87.5	=SUM('Raw Data (EAM)'!Y80:Y89)/SUM('Population (EAM)'!U80:U89)*10^5	=SUM('Raw Data (EAM)'!Y90:Y99)/SUM('Population (EAM)'!U90:U99)*10^5
92.5	=SUM('Raw Data (EAM)'!Z85:Z94)/SUM('Population (EAM)'!V85:V94)*10^5	=SUM('Raw Data (EAM)'!Z95:Z104)/SUM('Population (EAM)'!V95:V104)*10^5
97.5	=SUM('Raw Data (EAM)'!AA90:AA99)/SUM('Population (EAM)'!W90:W99)*10^5	=SUM('Raw Data (EAM)'!AA100:AA109)/SUM('Population (EAM)'!W100:W109)*10^5
102.5	=SUM('Raw Data (EAM)'!AB95:AB104)/SUM('Population (EAM)'!X95:X104)*10^5	=SUM('Raw Data (EAM)'!AB105:AB114)/SUM('Population (EAM)'!X105:X114)*10^5

Decades (EA)

Fig. 20: MORTALITY TEMPLATE: Age-specific mortality curves by decade of birth.

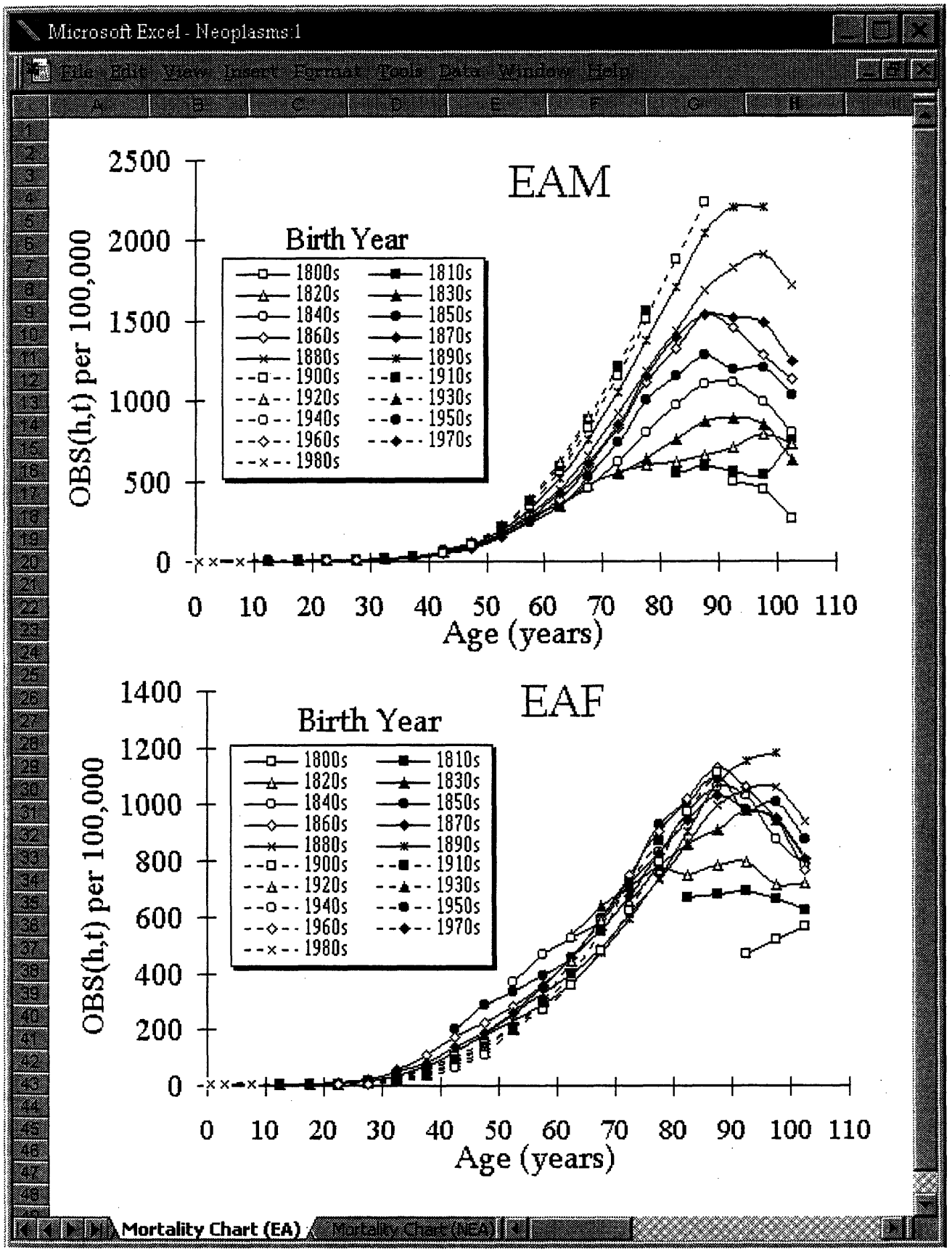

Microsoft Excel - Neoplasms1
File Edit View Insert Format Tools Data Window Help
EAM
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
OBS(h,t) per 100,000
2500
2000
1500
1000
500
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)
EAF
Birth Year
1800s
1810s
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
OBS(h,t) per 100,000
1400
1200
1000
800
600
400
200
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)
Mortality Chart (EA)
Mortality Chart (NEA)

Fig. 21: MORTALITY TEMPLATE: Age-specific cancer mortality curves by decade of birth (up to age 35).

Reveals occurrence of childhood cancers.

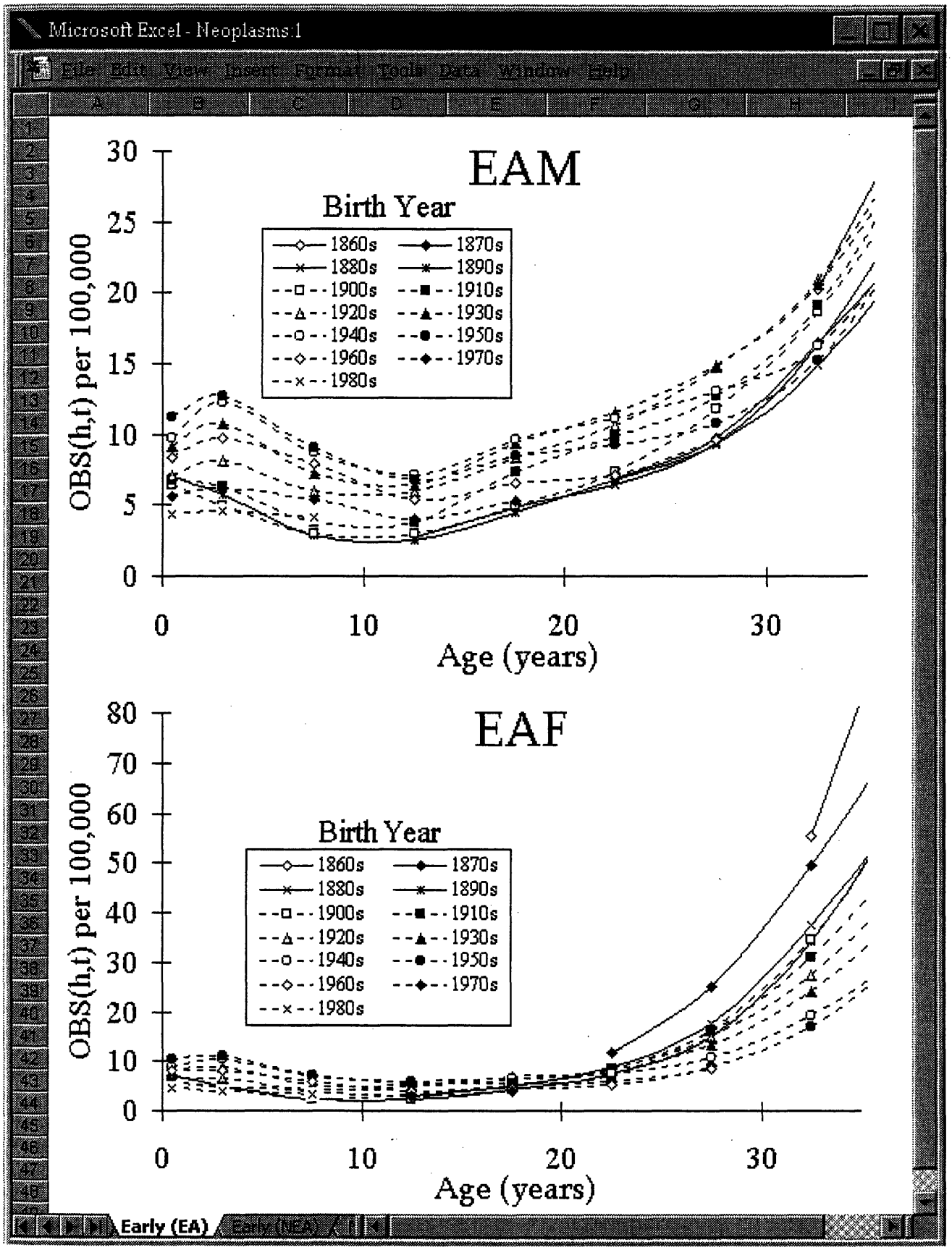

Microsoft Excel - Neoplasms1
File Edit View Insert Format Tools Data Window Help
EAM
Birth Year
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
OBS(h,t) per 100,000
Age (years)
EAF
Birth Year
OBS(h,t) per 100,000
Age (years)
Early (EA)
Early (NEA)

Fig. 22: MORTALITY TEMPLATE: Cancer mortality trends for 52.5, 62.5, and 72.5 year olds as a function of birthyear cohort.

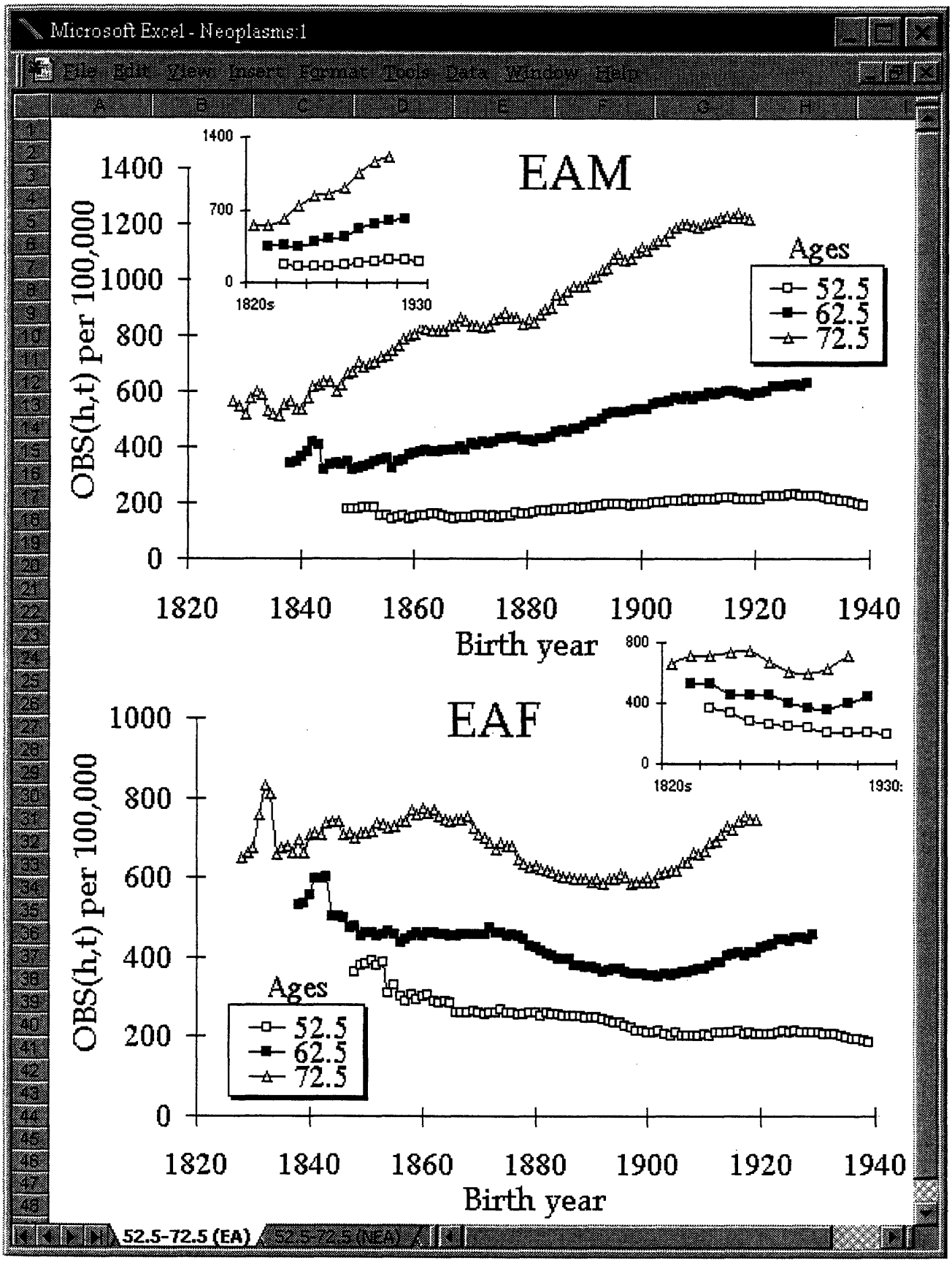

Microsoft Excel - Neoplasms:1
File Edit View Insert Format Tools Data Window Help
EAM
OBS(h,t) per 100,000
1400
1200
1000
800
600
400
200
0
1820 1840 1860 1880 1900 1920 1940
Birth year
Ages
52.5
62.5
72.5
1400
700
0
1820s
1930
EAF
OBS(h,t) per 100,000
1000
800
600
400
200
0
1820 1840 1860 1880 1900 1920 1940
Birth year
Ages
52.5
62.5
72.5
800
400
0
1820s
1930:
52.5-72.5 (EA)
52.5-72.5 (NEA)

Fig. 23: Mortality trends for 52.5, 62.5, and 72.5 year olds as a function of birthyear cohort for diabetes

Plot reveals drop-off in mortality rates by diabetes corresponding to the reporting year of 1949. Deaths by heart disease among diabetics were reported to be due to diabetes prior to 1949, but due to heart disease since 1949.

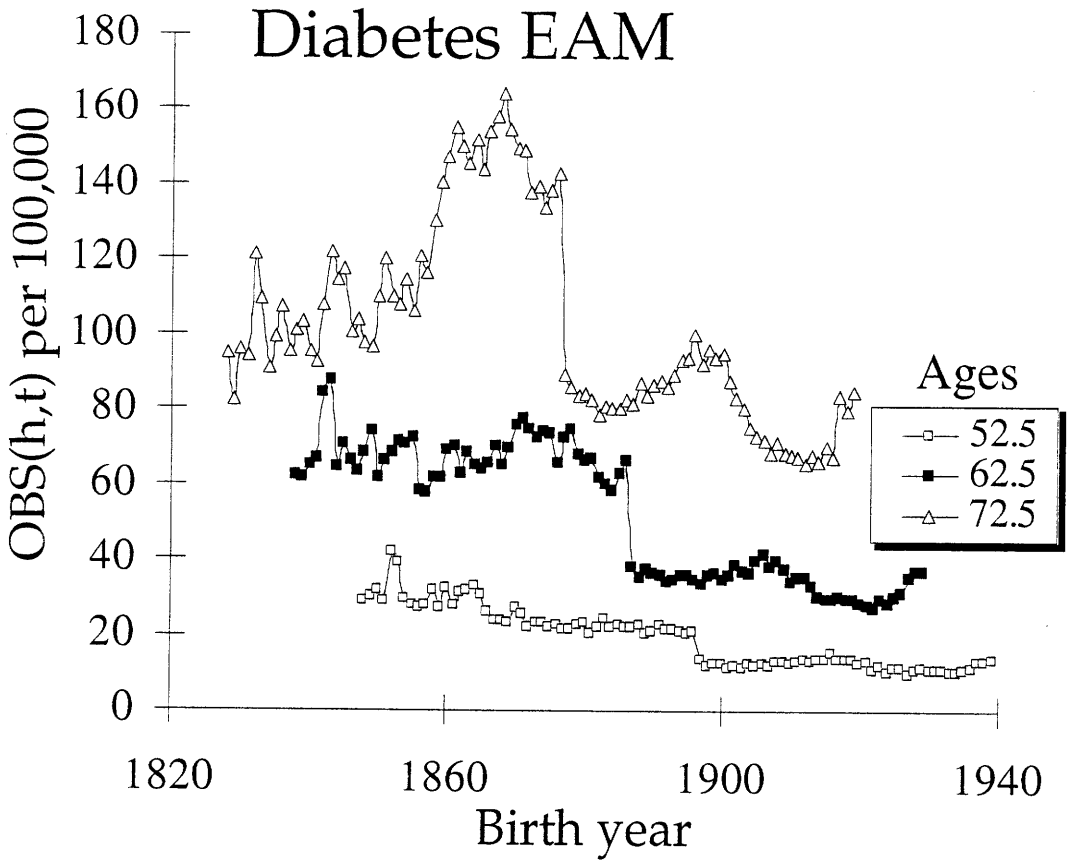
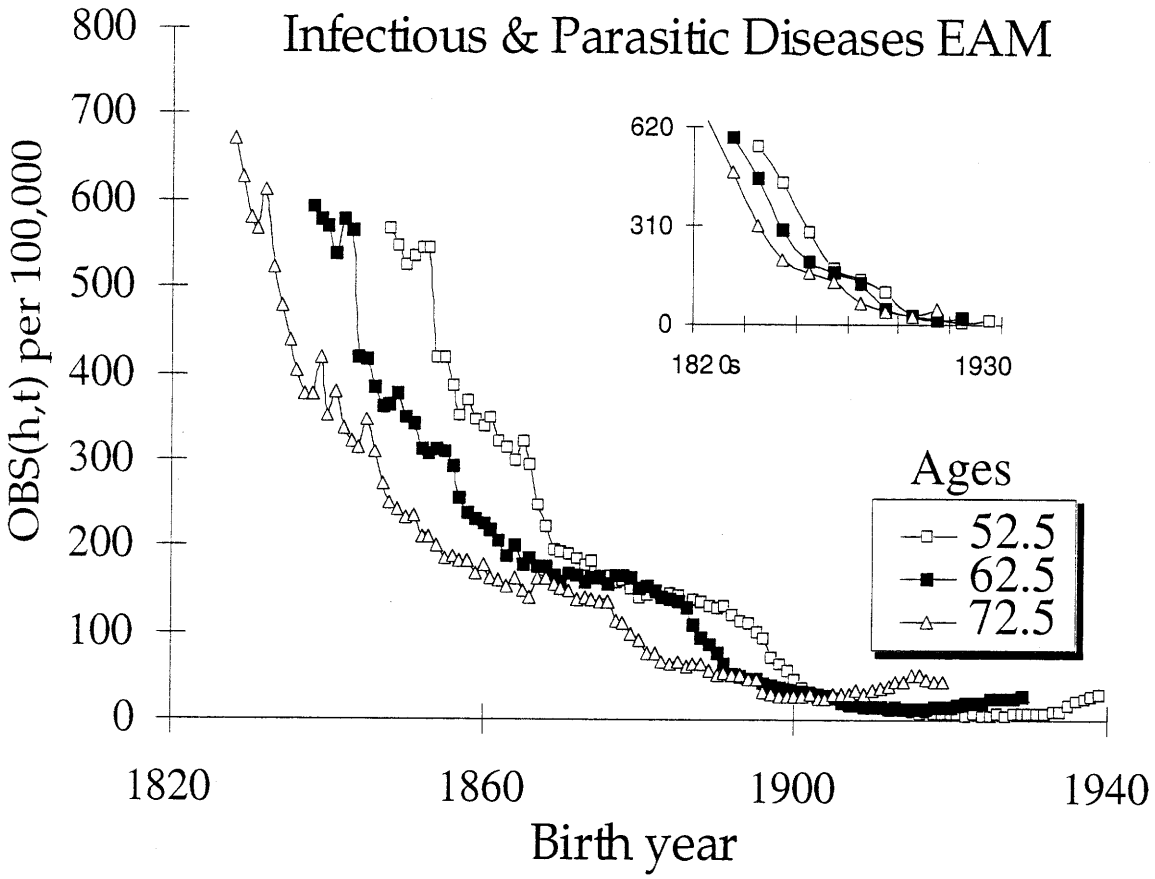
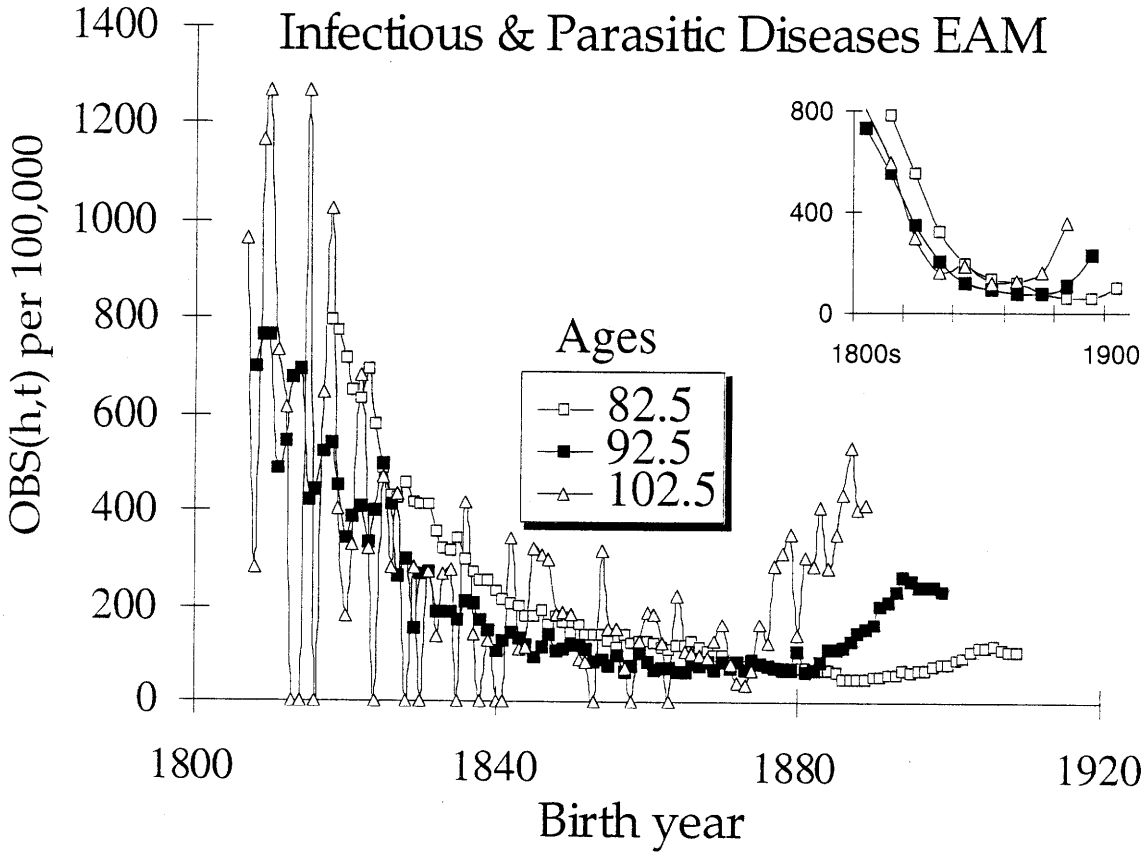

Diabetes EAM
OBS(h,t) per 100,000
180
160
140
120
100
80
60
40
20
0
1820
1860
1900
1940
Birth year
Ages
52.5
62.5
72.5

Fig. 24: Mortality trends for age groups > 50 years as a function of birthyear cohort for infectious and parasitic diseases

Plots reveal: drop-off in mortality rates by infectious diseases corresponding to improving sanitation methods during the last two centuries, drop-off in mortality rates by infectious diseases after the advent of antibiotics (drop after mortality rates had reached a plateau), and the increase in mortality among the elderly during the more recent years.


Infectious & Parasitic Diseases EAM
OBS(h,t) per 100,000
800
700
600
500
400
300
200
100
0
620
310
0
1820s
1930
Ages
52.5
62.5
72.5
1820
1860
1900
1940
Birth year


Infectious & Parasitic Diseases EAM
OBS(h,t) per 100,000
1400
1200
1000
800
600
400
200
0
800
400
0
1800s
1900
Ages
82.5
92.5
102.5
1800
1840
1880
1920
Birth year

Alternately, mortality data by infectious diseases plotted in this method (Figure 24) reveals the time point of the advent of antibiotics (the decrease in mortality after mortality rates had appeared to plateau). Furthermore, it reveals that prior to the advent of antibiotics, mortality by infectious diseases had dramatically decreased, presumably due to improvements in sanitation systems. Curiously, they also reveal that mortality by infectious diseases has actually increased among the elderly in the last decade, due to unknown reasons.

3.2.3.1 Colon Cancer Mortality Rates

Figures 25 and 26 summarize the age-specific intestinal cancer mortality records for the birth years between 1840 and 1930 for European and Non-European-Americans respectively. Intestinal cancer records are available since 1930, as opposed to 1958 for colon cancer when specific diagnoses became available. Intestinal cancer data is used herein to approximate colon cancer deaths; deaths by cancer of the small intestine represented only 3% of the total number of deaths from intestinal cancer in the period during which colon cancer was specifically recorded (U.S. Department of Health and Human Services (Vital Statistics of the United States, 1958-1992).

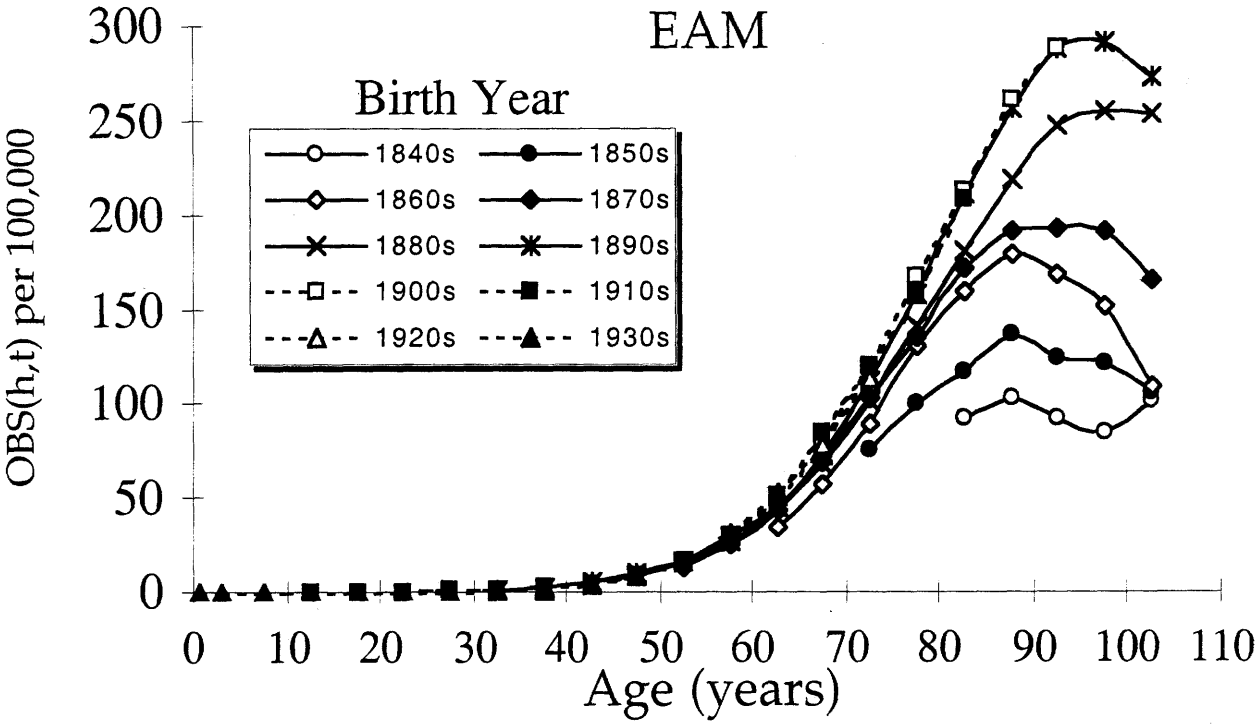
3.2.3.2 Lung Cancer Mortality Rates

Figures 27 and 28 summarize the age-specific intestinal cancer mortality records for the birth years between 1820 and 1990 for European and Non-European-Americans respectively. One notes that there are no apparent differences between populations of European or non-European, predominantly African, ancestry with regard to the historical record of age-specific lung cancer mortality rates.

Fig. 25: Intestinal cancer age- and birthyear- specific mortality curves

(EAM - European-American males, EAF - European-American females)

(Data recorded 1930-1992)


EAM
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS(h,t) per 100,000
300
250
200
150
100
50
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

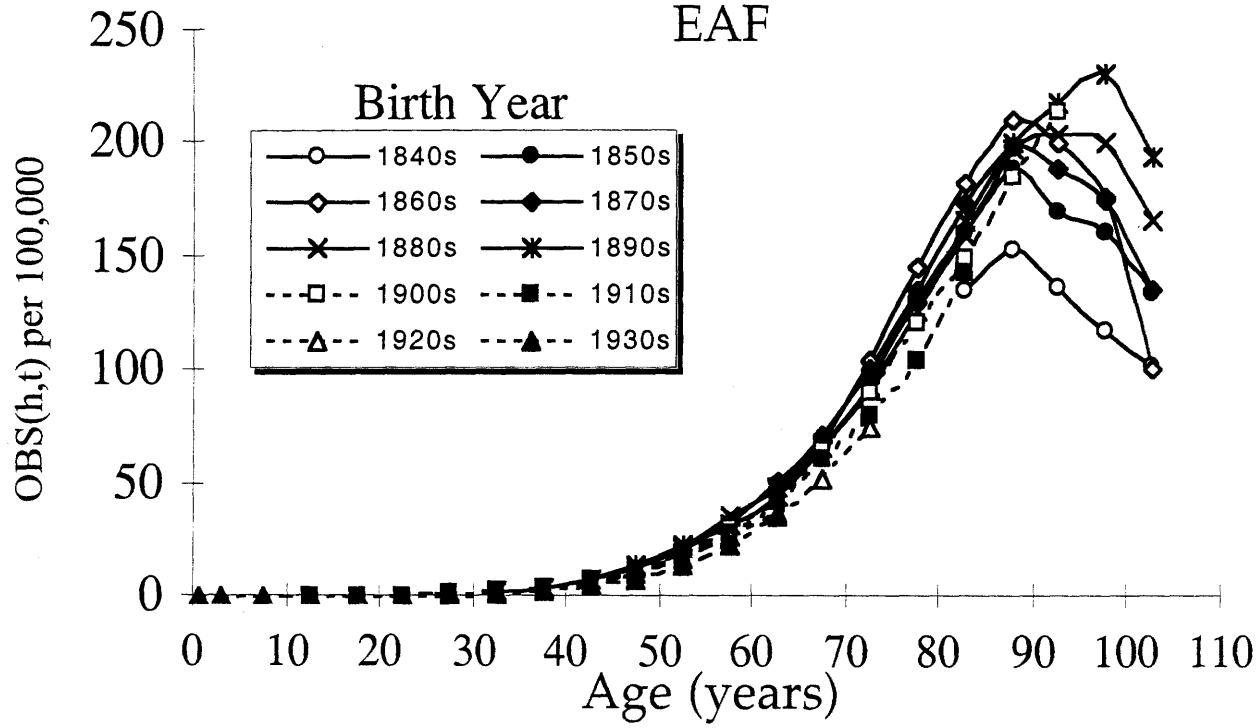
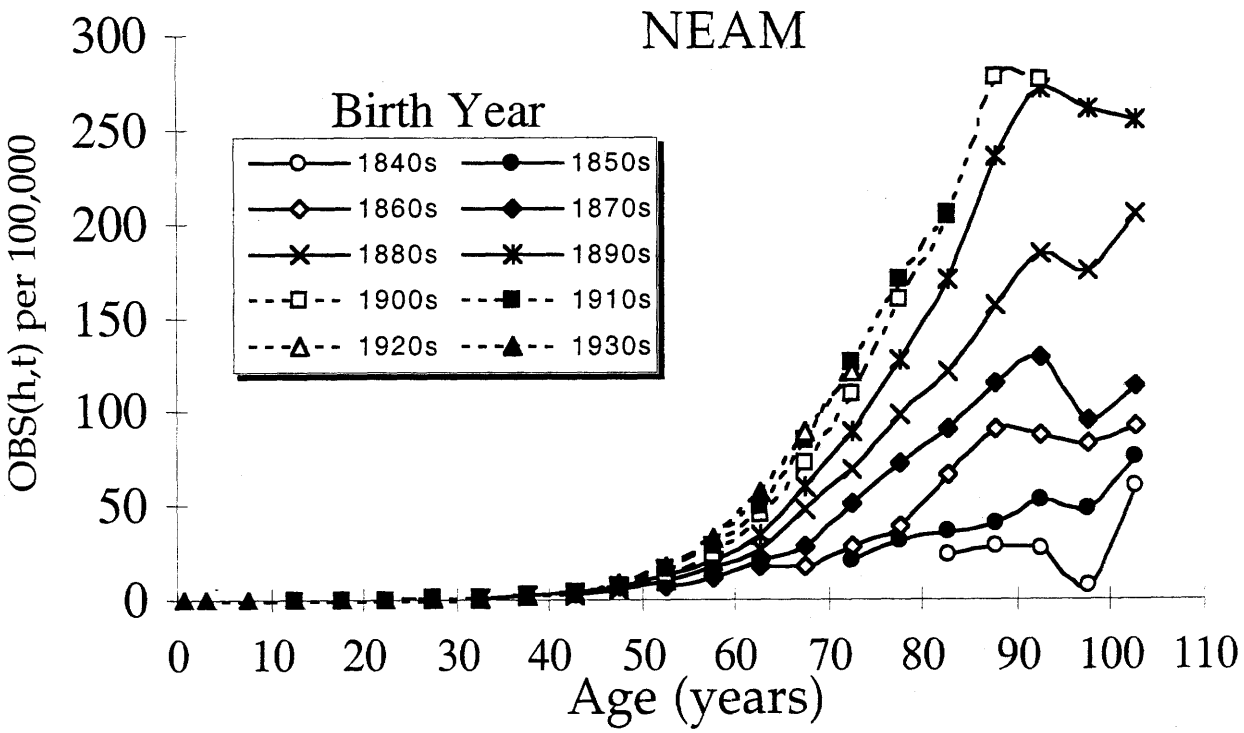

EAF
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS(h,t) per 100,000
250
200
150
100
50
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

Fig. 26: Intestinal cancer age- and birthyear- specific mortality curves

(NEAM – Non-European-American males, NEAF – Non-European-American females; primarily of African-American descent)

(Data recorded 1930-1992)


NEAM
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS(h,t) per 100,000
300
250
200
150
100
50
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

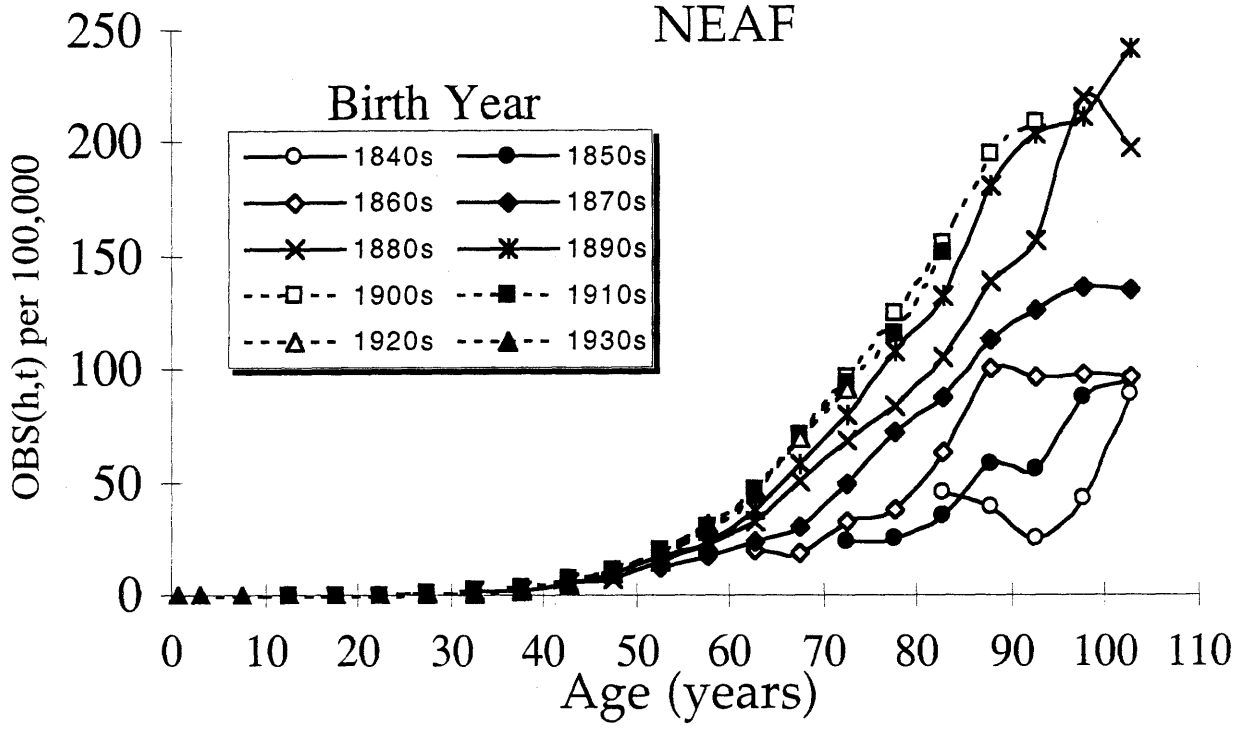
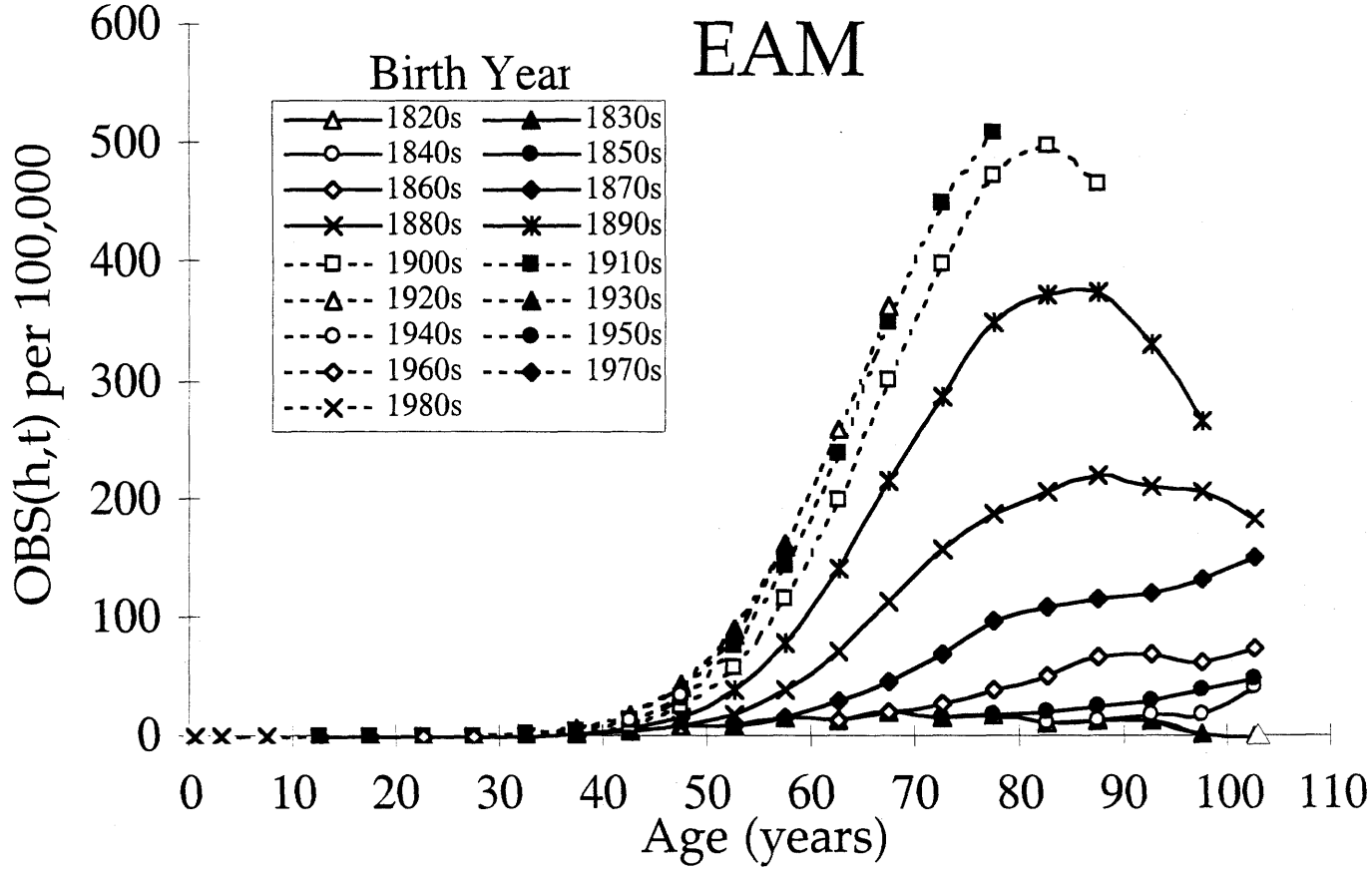

NEAF
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS(h,t) per 100,000
250
200
150
100
50
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

Fig. 27: Lung cancer age- and birthyear- specific mortality curves

(EAM - European American males, EAF - European American females)

Date recorded (1930-1992)


EAM
Birth Year
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
OBS(h,t) per 100,000
600
500
400
300
200
100
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

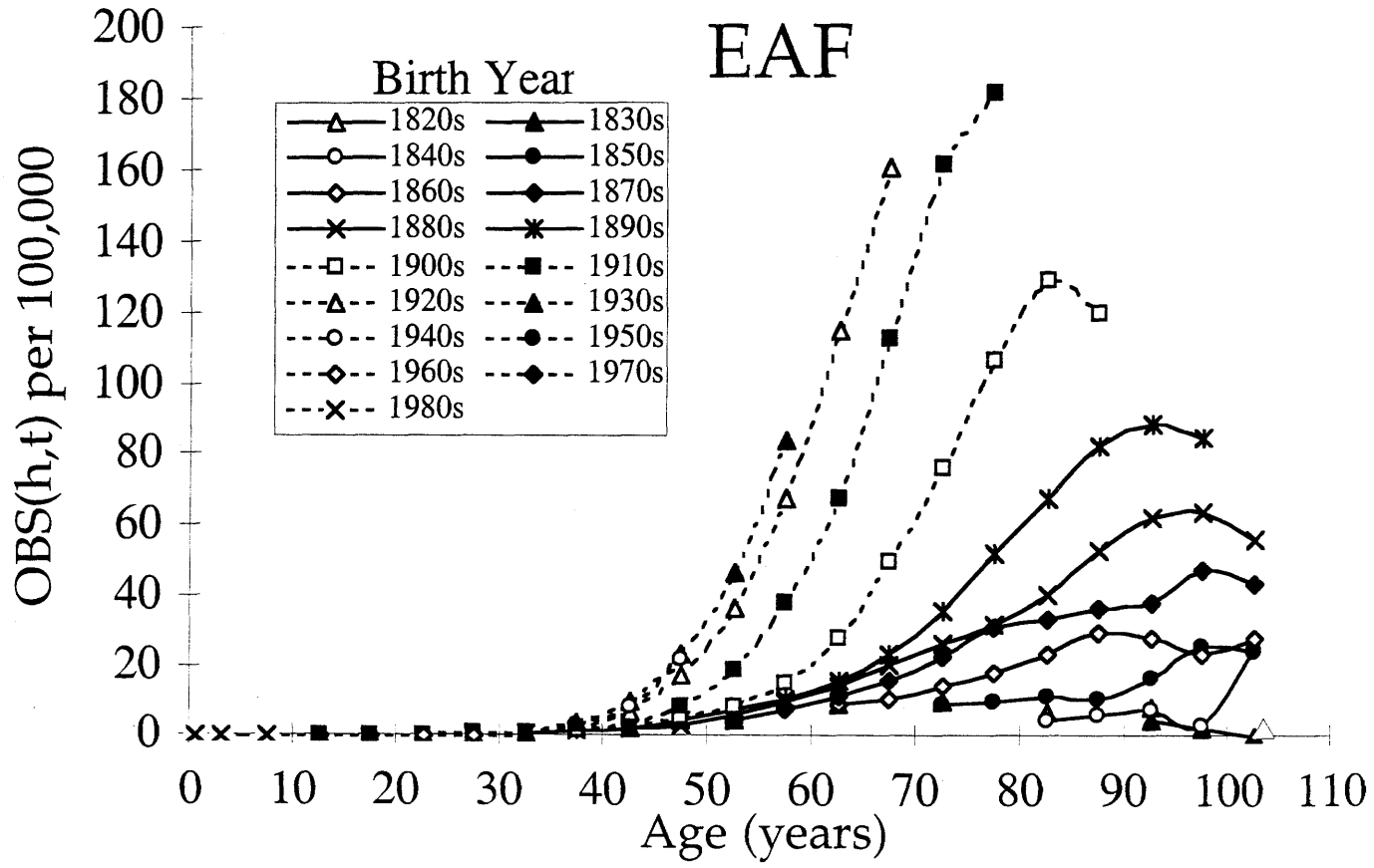
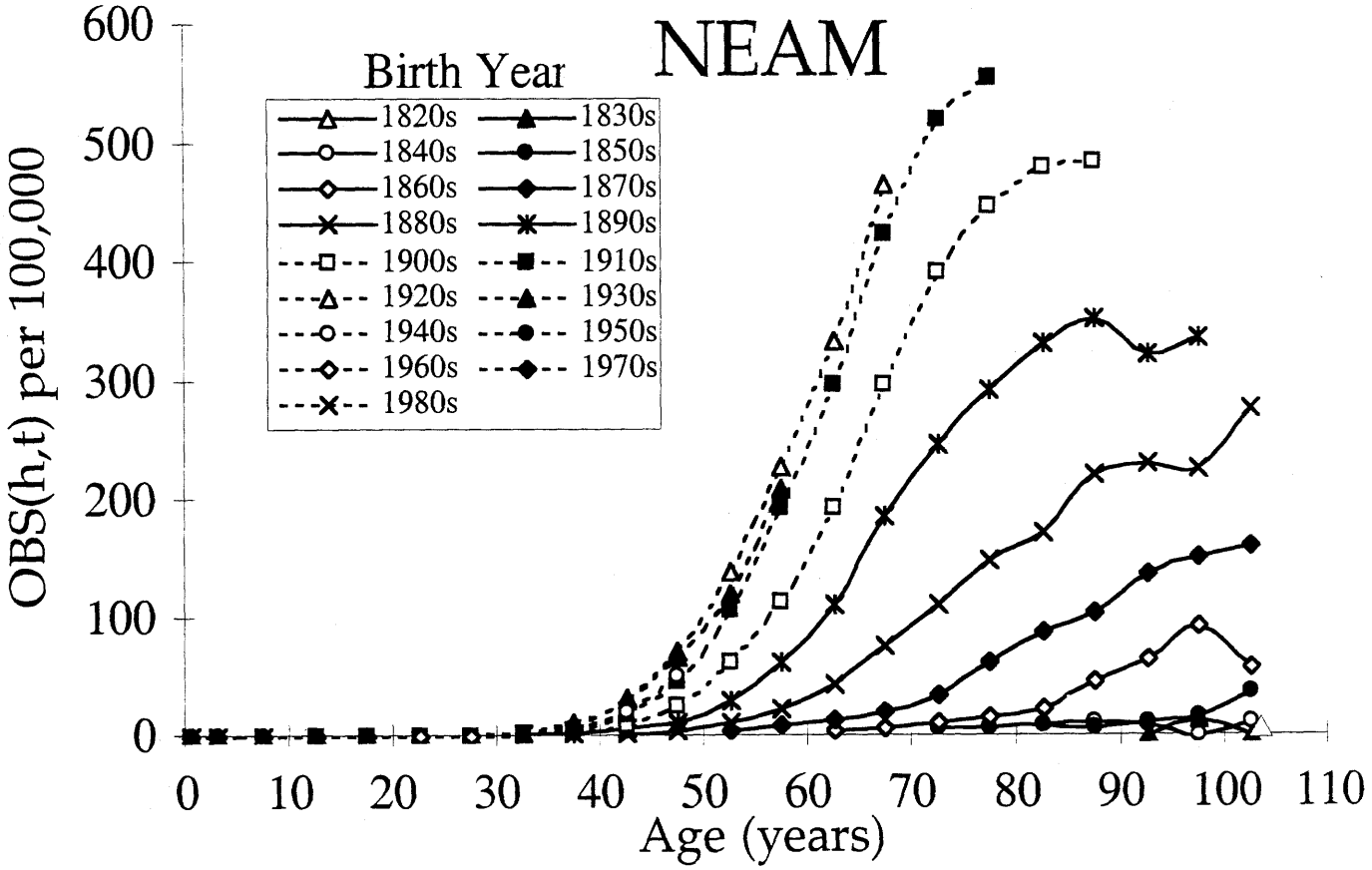
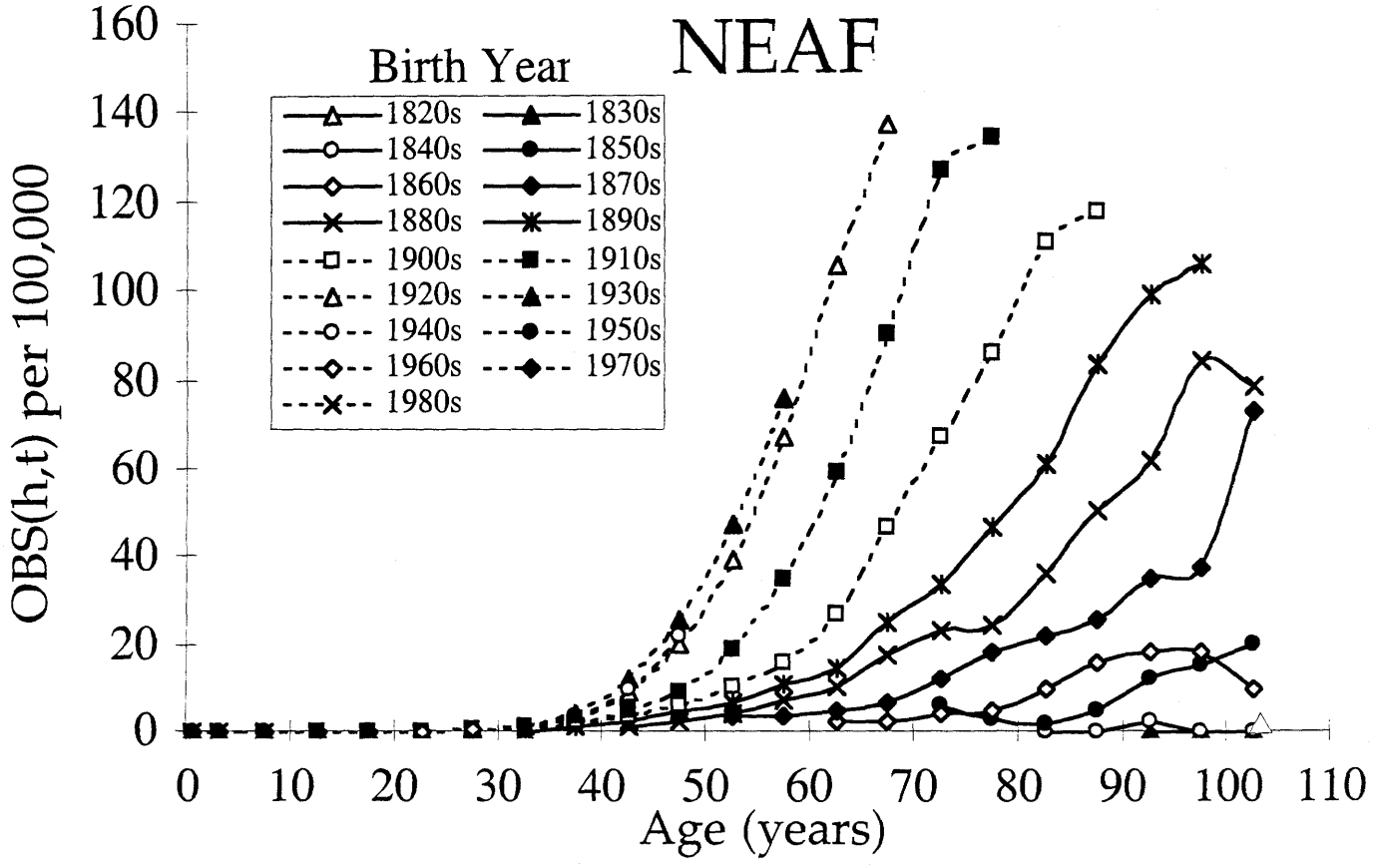

EAF
Birth Year
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
OBS(h,t) per 100,000
200
180
160
140
120
100
80
60
40
20
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

Fig. 28: Lung cancer age- and birthyear- specific mortality curves

(NEAM - Non-European American males, NEAF - Non-European American females; primarily of African descent)

Data recorded (1930-1992)


NEAM
Birth Year
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
OBS(h,t) per 100,000
0
100
200
300
400
500
600
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)


NEAF
Birth Year
1820s
1830s
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
1960s
1970s
1980s
OBS(h,t) per 100,000
0
20
40
60
80
100
120
140
160
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

3.2.4 Survival Rate Data (Cancer)

For each birth year cohort h and each age t there is an associated relative 5-year survival rate for cancer, S(h,t). S(h,t) represents the probability of surviving those causes linked to that cancer. **(Eq.2)**

$$S(h,t) = \frac{\text{(recorded cancer survivors at age t+5, diagnosed at age t)}}{\text{(recorded diagnoses of cancer at age t) x (survival rate for all forms of death, age t + 5)}}$$

Relative survival rates correct for the fraction of individuals diagnosed with cancer who may have died of an unrelated form of death within the 5-year period after diagnosis.

Eisenberg et al (1968) have summarized the age-specific relative survival cancer rates for 1935 to 1959 within the state of Connecticut for both males and females up to ages 65-74, and for ages greater than 75. The NCI Monograph No. 6 (1961) summarized similar relative survival rates for 1950 to 1957 for both males and females, including both a larger set of hospital registries, and relative survival rates for untreated individuals. The Cancer Patient Survival Report Number 5 (1976) extended this work to the period of 1950 to 1972, including survival rates for patients of both European and African descent. Ries et al (1983) similarly reports the relative survival rates for the period of 1973 to 1975 by age, gender, and race. Last, the SEER Cancer Statistics Reviews (1993, 1997, 1999) have recorded the 5-year relative survival rates for 1983 through 1991.

Reported survival rates do not account for those deaths of individuals first diagnosed with cancer at the time of death. The percentage of 'incidences at autopsy' is 1-2% for the 1990s [personal communication, L.A.G. Ries, SEER]. For diagnostic years 1935-79 the percentage of

'incidences at autopsy' for the state of Connecticut was generally 1-3%, but was shown to increase as a function of age (Heston et al, 1986).

3.2.4.1 Survival Data – Colon Cancer

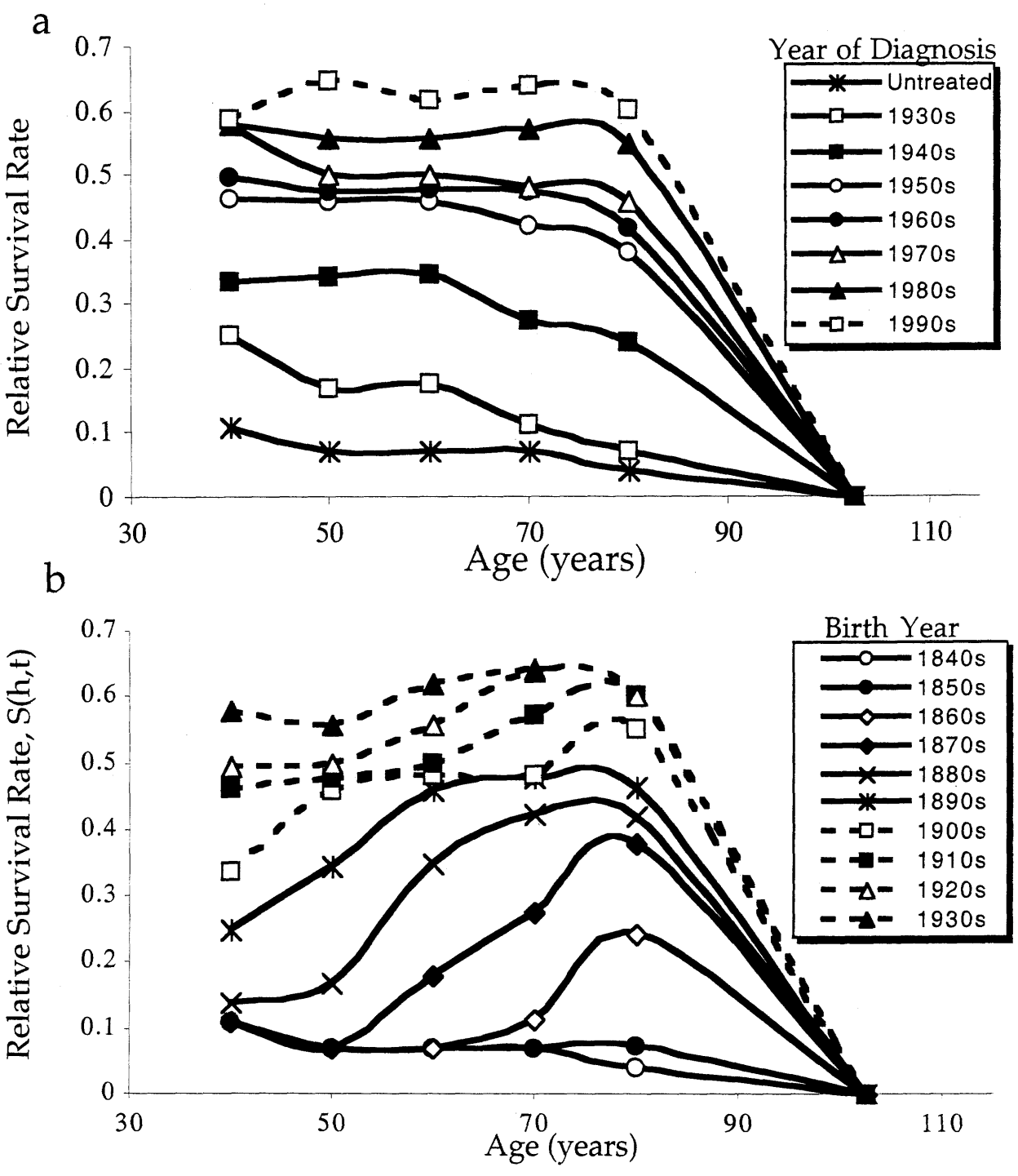
Aside from the survival rate reports listed in Section 3.2.4, Beart et al (1995) have reported the age-specific relative survival rates for 1983, although gender and race were not specified. Beart's overall survival estimates for the early 1980s were about 10% lower than as reported by SEER (1999). Consequently for the 1980s, SEER's reported survival rates were decreased by 5% to represent the average reported survival rates of SEER and Beart et al (1995).

Colon cancer survival rates are approximately constant between ages 40 and 75 in recent decades (Ries et al, 1983; SEER 1993, 1997, 1999). However, Beart et al (1995) extended the survival data to 80 years and found survival rates decreased significantly from 70-79 and >80 years of age. Survival rates appear to decrease even further in extreme old age when colon cancer is more often detected in an advanced stage. In persons over 80 years of age, only 2.4% of all colon tumors were treated by surgery and chemotherapy, compared to 26.3% for persons less than 50 year olds (Beart et al, 1995). As an estimate, survival rate for untreated tumors of 75+ year olds were used to estimate survival rate of centenarians, 3-4% (NCI Monograph No. 6, 1961).

Figure 29 illustrates survival rates for European-American females born between the 1840s and 1930s. The data recorded by year of diagnosis in Figure 29 are converted into age-specific relative survival values by year of birth. Where values were unknown, estimates were interpolated. Survival rates reported for the age ranges "under 45" and "above 75", are plotted at ages 40 and 80, respectively, as these are approximately the average ages of individuals dying of

Fig. 29: Survival rates - Colon Cancer

Age-specific relative survival rates for colon cancer among European-American females (a) by year of diagnosis and (b) by year of birth.


a
Relative Survival Rate
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
30
50
70
90
110
Age (years)
Year of Diagnosis
Untreated
1930s
1940s
1950s
1960s
1970s
1980s
1990s
b
Relative Survival Rate, S(h,t)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
30
50
70
90
110
Age (years)
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s

colon cancer in these age groups. S(h,t) increases steadily with historical time which creates a steady age-specific increase in S(h,t) for any particular birth year cohort, but which still decreases markedly in extreme old age.

Table 1 summarizes the relative survival rates by year of diagnosis, using averages for those years for which more than one value was available. Averages were weighted according to the number of patients examined in each study. To estimate relative survival rates for Non-European-Americans, we used reports for African-American survival rate data set for the 1950s through the 1990s, as African-Americans comprised more than 75% of the Non-European population. Estimates for the 1940s and 1930s cohorts of Non-Europeans were interpolated assuming that the change in the survival rate for European-Americans was proportional to the change in the survival rate of Non-European-Americans during this period. No estimates were allowed to drop below the reported survival rates for untreated individuals.

Table 1: Summary of the relative survival rates by year of diagnosis

1990s Ages:	0-44	45-54	55-64	65-74	75+	100+
EAM	0.58	0.62	0.65	0.66	0.59	0.03
EAF	0.59	0.65	0.62	0.64	0.60	0.04
NEAM	0.51	0.54	0.55	0.52	0.45	0.03
NEAF	0.55	0.53	0.56	0.53	0.46	0.04
1980s Ages:	**0-44**	**45-54**	**55-64**	**65-74**	**75+**	**100+**
EAM	0.49	0.59	0.59	0.60	0.57	0.03
EAF	0.58	0.56	0.56	0.57	0.55	0.04
NEAM	0.44	0.49	0.49	0.47	0.37	0.03
NEAF	0.52	0.54	0.53	0.43	0.41	0.04
1970s Ages:	**0-44**	**45-54**	**55-64**	**65-74**	**75+**	**100+**
EAM	0.47	0.48	0.48	0.48	0.44	0.03
EAF	0.58	0.50	0.50	0.48	0.46	0.04
NEAM	0.42	0.46	0.45	0.38	0.32	0.03
NEAF	0.53	0.50	0.45	0.50	0.37	0.04

1960s	Ages:	0-44	45-54	55-64	65-74	75+	100+
EAM		0.50	0.45	0.45	0.44	0.37	0.03
EAF		0.50	0.48	0.48	0.47	0.42	0.04
NEAM		0.29	0.42	0.31	0.29	0.25	0.03
NEAF		0.36	0.46	0.38	0.30	0.34	0.04

1950s	Ages:	0-44	45-54	55-64	65-74	75+	100+
EAM		0.42	0.46	0.40	0.38	0.32	0.03
EAF		0.46	0.46	0.46	0.42	0.38	0.04
NEAM		0.28	0.37	0.25	0.32	0.18	0.03
NEAF		0.44	0.36	0.33	0.24	0.15	0.04

1940s	Ages:	0-44	45-54	55-64	65-74	75+	100+
EAM		0.27	0.33	0.29	0.21	0.17	0.03
EAF		0.34	0.34	0.35	0.28	0.24	0.04
NEAM		0.18	0.26	0.18	0.18	0.09	0.03
NEAF		0.30	0.27	0.25	0.16	0.10	0.04

1930s	Ages:	0-44	45-54	55-64	65-74	75+	100+
EAM		0.30	0.27	0.20	0.09	0.00	0.03
EAF		0.25	0.17	0.18	0.11	0.07	0.04
NEAM		0.20	0.22	0.12	0.07	0.04	0.03
NEAF		0.22	0.13	0.13	0.07	0.04	0.04

Untreated*	Ages:	0-44	45-54	55-64	65-74	75+
Males		0.00	0.10	0.11	0.07	0.03
Females		0.11	0.07	0.07	0.07	0.04

* Reported as Other and Untreated (NCI Monograph No. 6, 1961).

3.2.4.2 Survival Data – Lung Cancer

Survival of lung cancer approaches 5% at fifteen years after diagnosis, with a 15% probability at five years. Survival is *pro tempore* approximated as zero.

3.2.5 Estimates of Error in Reported Data

It is obvious that the numerator defining OBS(h,t) in Equation 1 will be affected by the probability that an actual cancer mortality is recorded as such. It is equally obvious that there are

no records of inadequate diagnosis *per se*. Improved estimates can be made by accounting for the number of deaths in a cohort without any adequate diagnosis as a function of age, which represents a relative marker of the potential accuracy of the data as a function of the reporting year and age group. For instance, in centenarians the percentage of deaths with vague diagnoses was about 20% in the 1930s for European-American males, decreasing to less than 5% by the 1950s.

Inspection of the historical record for the number of deaths with vague or unrecorded diagnoses for all ages, genders, and ethnic groups, for each birth year cohort analyzed, creates a matrix for each demographic group defining an estimate of the probability of accurately recording the cause of death as the function R(h,t).

(Eq. 3)

$$R(h,t) = \frac{\text{recorded deaths from specified causes from birth cohort h at age t}}{\text{all recorded deaths from birth cohort h at age t}}$$

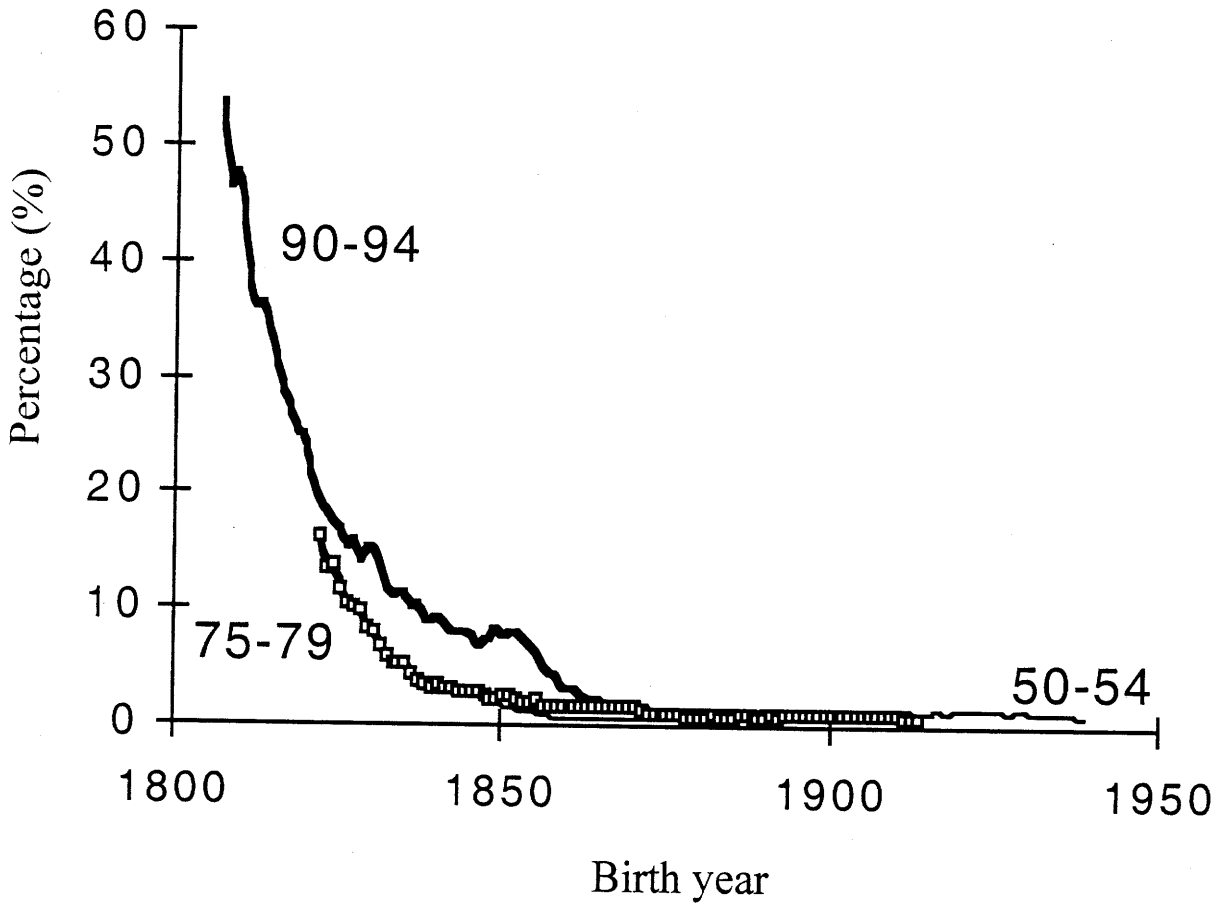
Figure 30 shows the percentage of all deaths with vague diagnoses plotted as a function of the birth year for several age groups of European-American males. The assumption here is that the proportion of cancer deaths among all deaths with unrecorded diagnoses is about the same as the proportion of cancers among all deaths with recorded diagnoses.

Application of this assumption still underestimates the true cancer mortality fraction. Since about 50% of present deaths are recorded as due to cardiovascular or cerebrovascular causes, small overestimates in these diagnoses would lead to large underestimates of mortality from any other specific disease.

Furthermore, a diagnosis of cancer may be in error, particularly if a detected mass were a secondary tumor from another organ. This kind of error has been addressed in a number of

Fig. 30: Percentage of all deaths with vague diagnoses

European-American males, ages 50-54, 75-79, and 90-94 as a function of their year of birth.


60
50
40
30
20
10
0
Percentage (%)
90-94
75-79
50-54
1800
1850
1900
1950
Birth year

studies in which pathological samples were reviewed. Colon cancer was actually found to be somewhat over-reported in death certificates primarily because of inclusion of a portion of rectal tumors (U.S. Department of Health and Human Services, 1982b). Lung cancer mortality rates were reported to be "fairly accurate" as tumors had both high detection and confirmation rates.

3.2.6 Mortality Data Adjusted for Historical and Age-Specific Survival Probability and Reporting Error – Colon and Lung Cancer

Combining available mortality data, survival data, and vague diagnoses data improve upon the estimates of actual occurrence rates of colon cancer. The amended data set is of sufficient accuracy to permit application of mathematical analyses, but exploration of the effect of errors in survival or reporting data on the estimation of parameters (Section 4.1.3) will be considered.

Figures 31 and 32 recast the data of Figures 27 and 28 using all of the estimates of S(h,t) and R(h,t) with OBS(h,t) to define a new function OBS*(h,t).

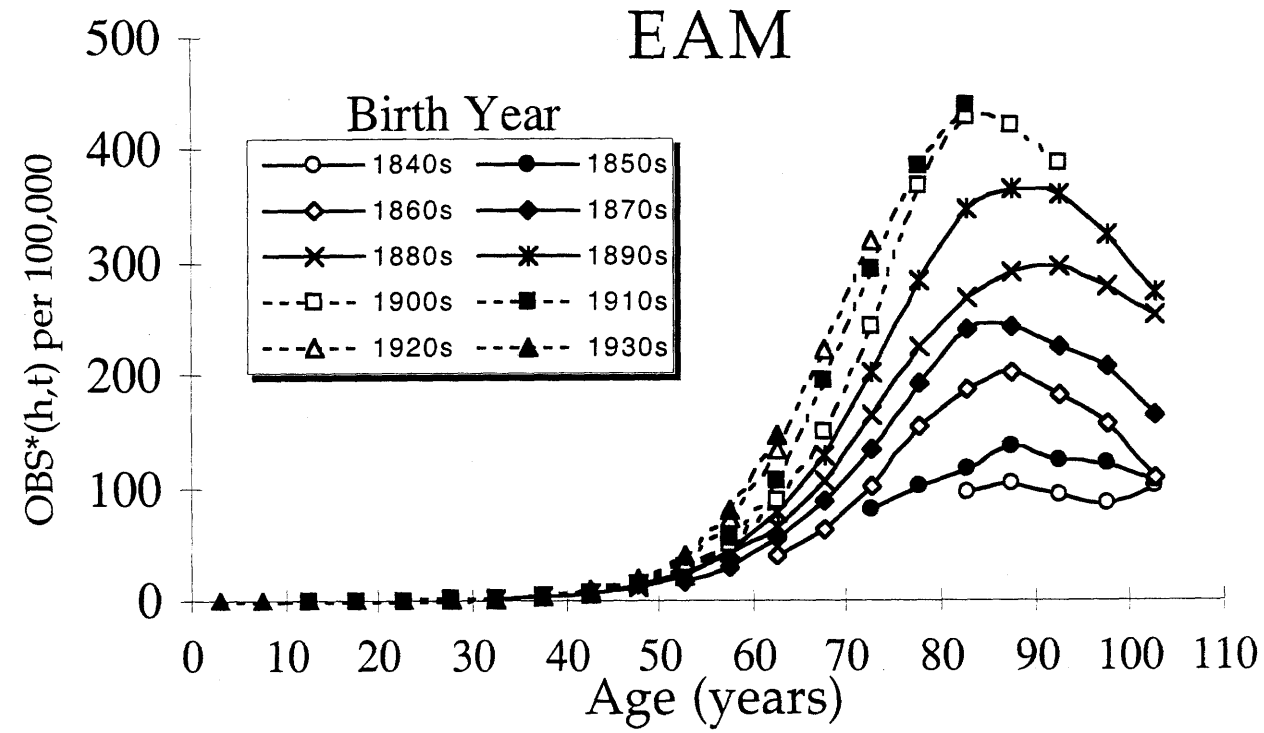
(Eq. 4) $$OBS^*(h,t) = OBS(h,t) \div [R(h,t)\,(1 - S(h,t))]$$

These figures are estimates of what colon cancer mortality rates would have been in a world with accurate diagnosis and recording but no therapy of any kind. In a sense it is a reconstruction of "incidence" data in a world with accurate diagnosis but without effective therapy.

In the case of lung cancer, since R(h,t) is approximately 1 for the years of death reported and S(h,t) is approximately zero, $OBS^*(h,t) \sim OBS(h,t)$.

Fig. 31: Colon cancer age- and birthyear- specific mortality curves adjusted for historical changes in underreporting and survival rates (European-Americans)

$$OBS^*(h,t) = OBS(h,t) \div [R(h,t)\,(1 - S(h,t))]$$


EAM
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS*(h,t) per 100,000
500
400
300
200
100
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

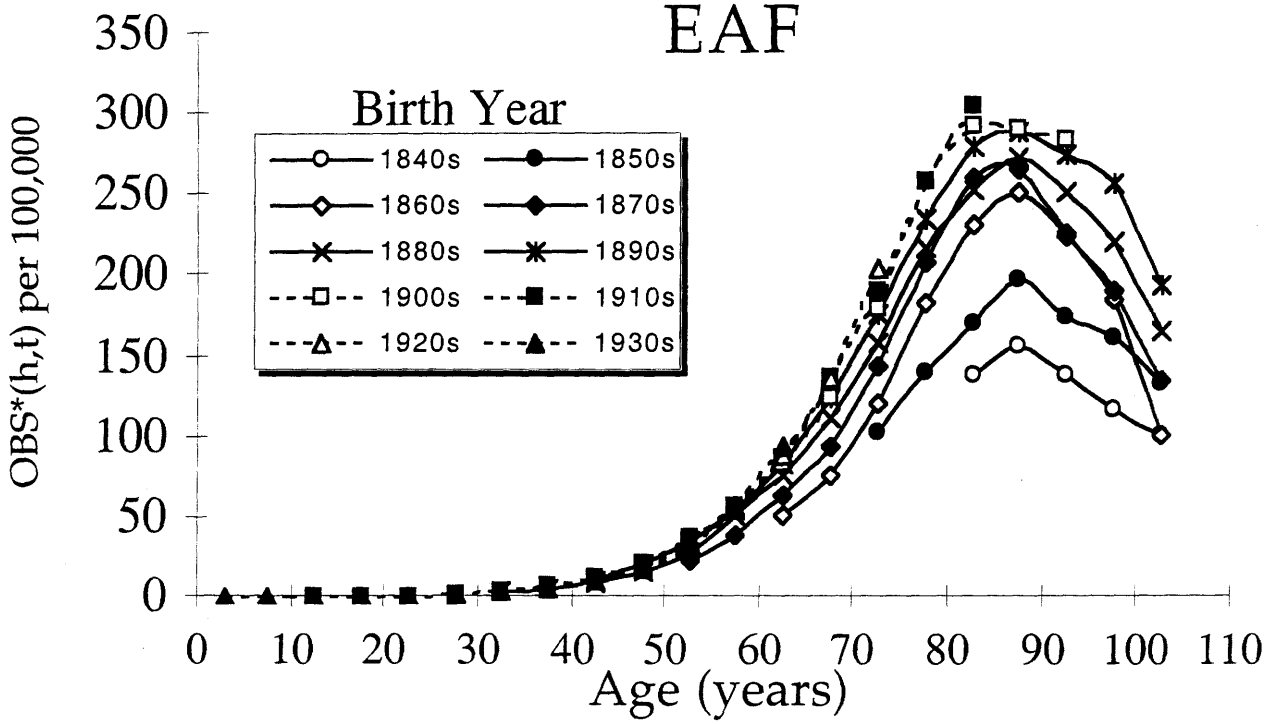
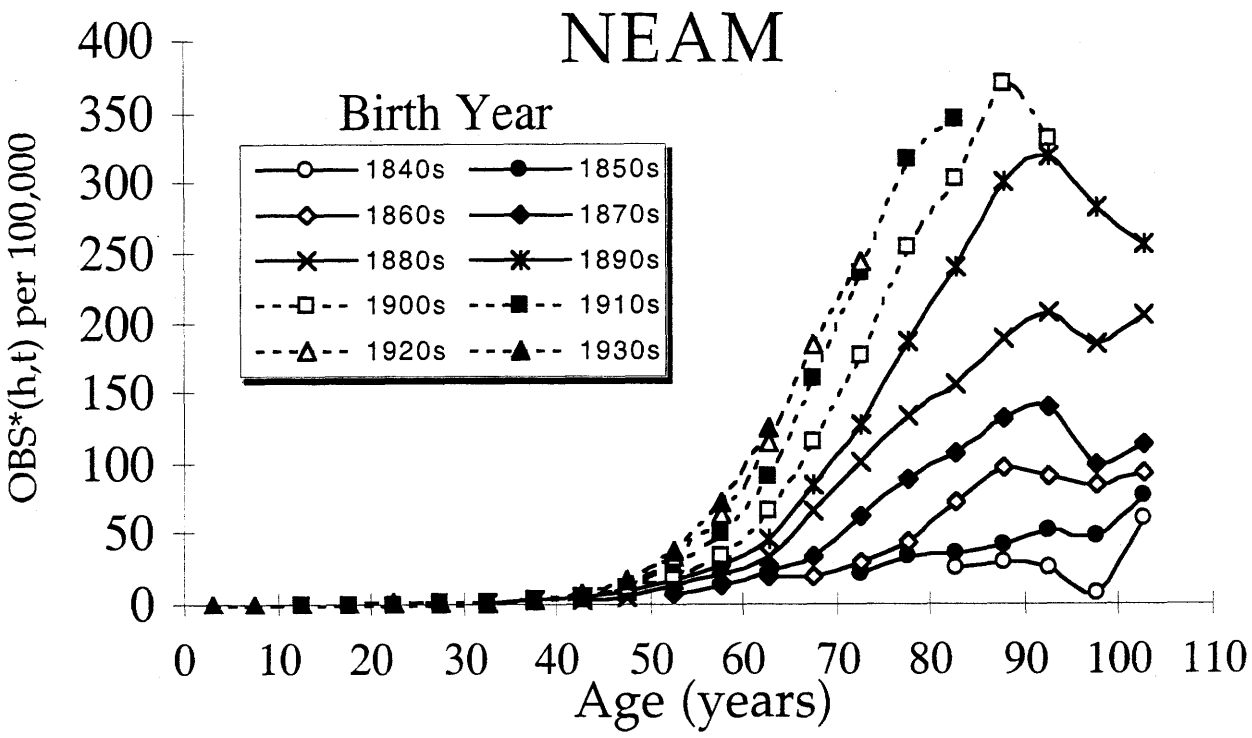
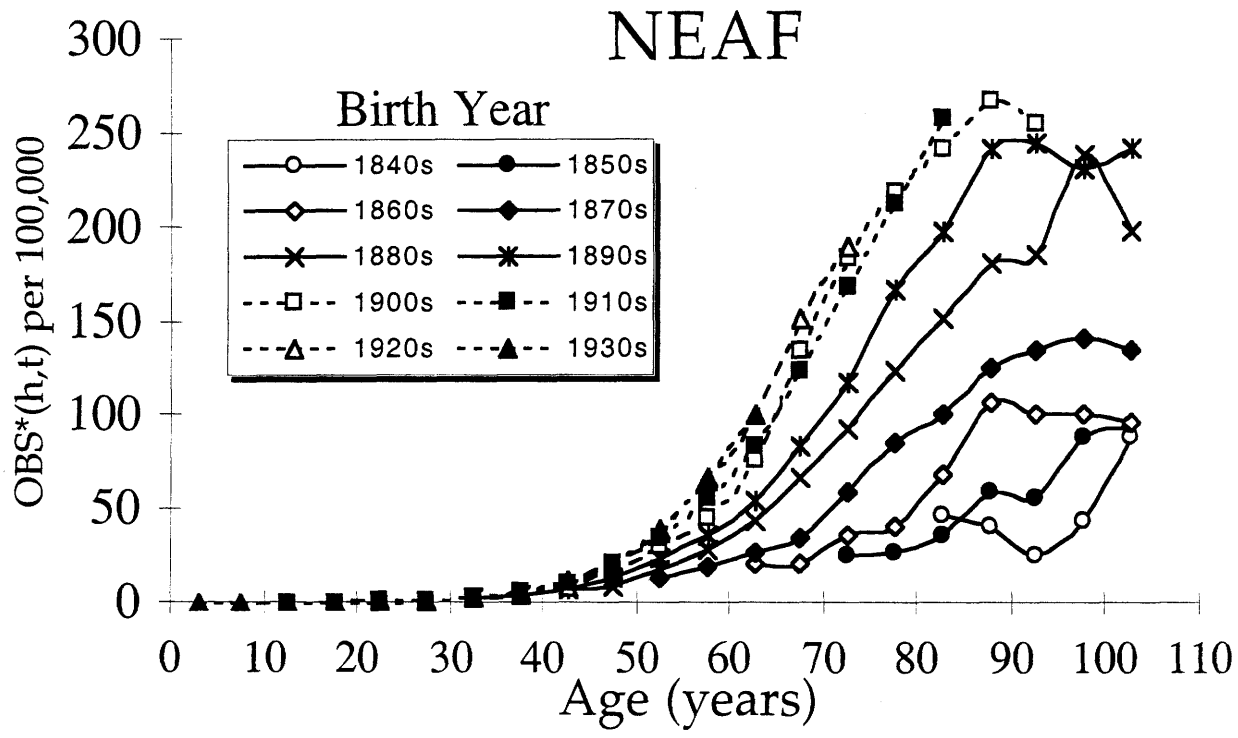

EAF
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS*(h,t) per 100,000
350
300
250
200
150
100
50
0
0 10 20 30 40 50 60 70 80 90 100 110
Age (years)

Fig. 32: Colon cancer age- and birth year- specific mortality curves adjusted for historical changes in underreporting and survival rates (Non-European Americans)

$$OBS^*(h,t) = OBS(h,t) \div [R(h,t)\,(1 - S(h,t))]$$


NEAM
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS*(h,t) per 100,000
400
350
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)


NEAF
Birth Year
1840s
1850s
1860s
1870s
1880s
1890s
1900s
1910s
1920s
1930s
OBS*(h,t) per 100,000
300
250
200
150
100
50
0
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

The importance of accounting for survival and reporting errors is illustrated by comparing the function OBS(1880s,t) for the EAM cohort of Figure 25 to the function OBS*(1880s,t) for the same cohort in Figure 31. In the former, OBS(h,t) 'appears' to reach a stable maximum plateau by age 90, but in the latter, OBS*(h,t) shows a clear maximum declining through age 102.5. A similar effect may be noted by comparing the NEAF cohort of OBS(1870s,t) to OBS*(1870s,t).

3.2.7 Prevalence of Cigarette Use.

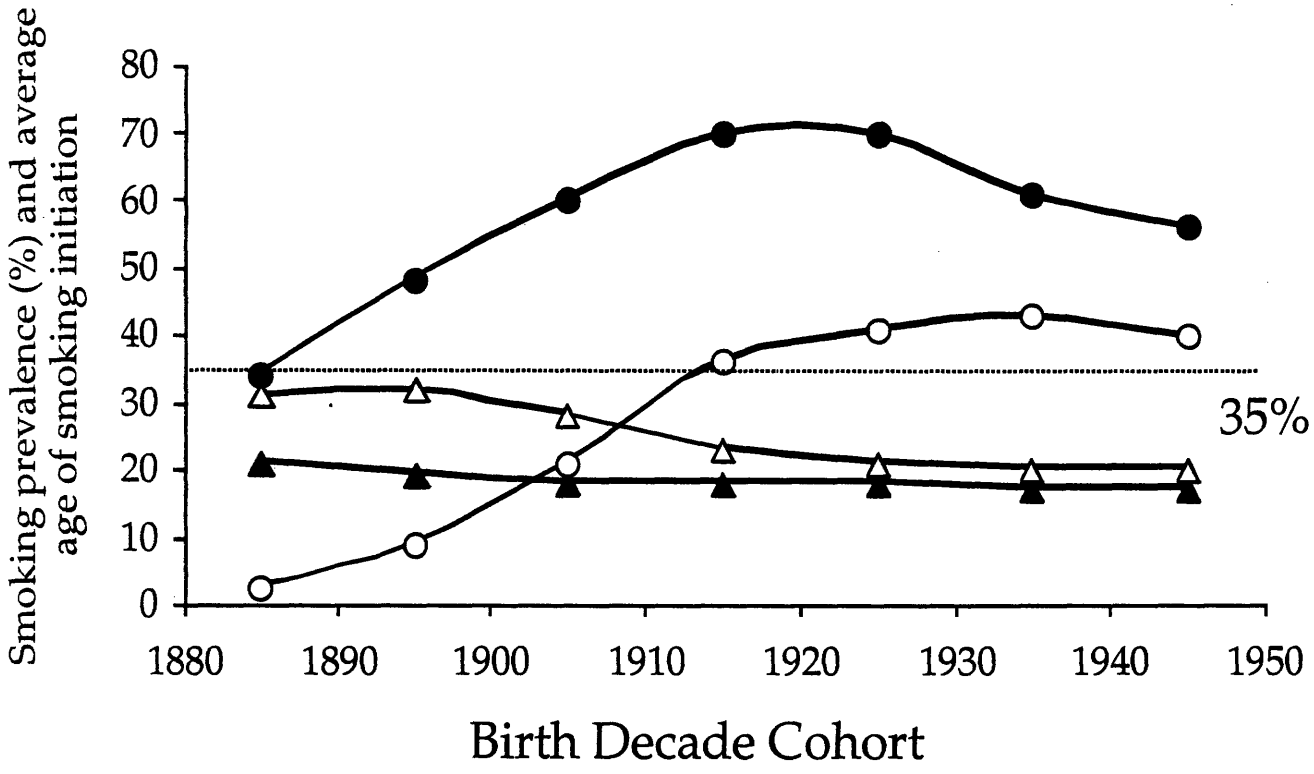
Prior to the 1880s, tobacco was predominantly chewed or smoked in pipes. Cigarettes became inexpensive and widely available after the 1880s with the invention of the automatic cigarette roller by Albert Bonsack (http://www.tobacco.org/History/Tobacco_History.html, "A Capsule History of Tobacco.")

Figure 33 shows the maximum cigarette-smoking fraction and the average age at which smoking started for each birth decade cohort since 1881-1890 as collected and organized by Harris (1983). Some smokers within each cohort began smoking very late in life or ceased smoking in middle age. Thus this maximum value is perforce an overestimate of the continuously smoking population. For those birth cohorts that began smoking in adolescence, lung cancer death rates rose rapidly at about age 50 (Figure 27), increased approximately linearly into old age, reached a distinct maximum and declined in extreme old age. In earlier birth cohorts that did not start using cigarettes until later in adulthood, lung cancer death rates also show a marked rise, but considerably later in life as may be seen by inspecting Figures 27 and 28.

Fig. 33: Smoking prevalence in the U.S.

Maximum US cigarette smoking fraction (circles) and average age at which smoking habit was adopted (triangles), organized by birth decade cohort. Male and female values are represented by closed and open markers respectively. The fraction of lifetime smokers is perforce somewhat lower than the maximum values.

(Derived from Harris, 1983)


Smoking prevalence (%) and average age of smoking initiation
80
70
60
50
40
30
20
10
0
1880
1890
1900
1910
1920
1930
1940
1950
35%
Birth Decade Cohort

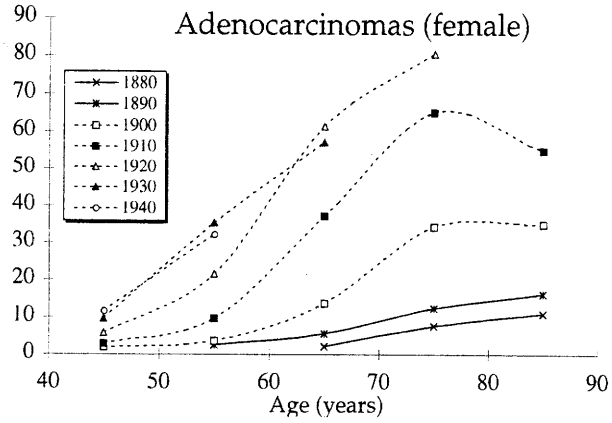
3.2.8 Histopathology of Lung Tumors.

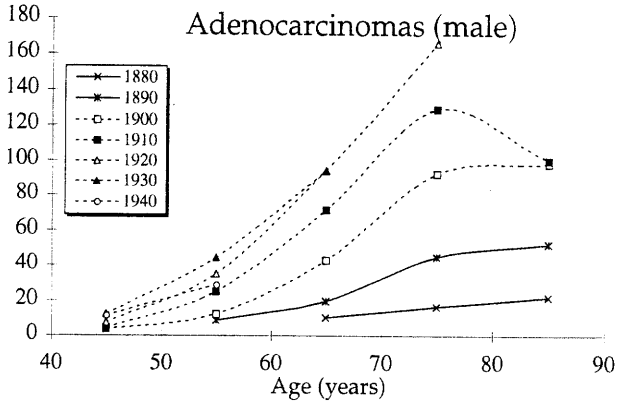
Zheng et al (1994) and Thun et al (1997) have independently examined the historical changes in lung cancer incidence by histologic type in the state of Connecticut. Age and gender-specific mortality rates for each of the three predominant histopathologic forms of lung cancer, squamous cell carcinoma, small cell carcinoma and adenocarcinoma are summarized in Figure 34. These data are reproduced from an earlier publication (Thun et al, 1997) and from additional data kindly put together and supplied by Dr. Thun. Additionally, bronchioloalveolar adenocarcinomas rates were shown to remain relatively constant (Zheng et al, 1994).

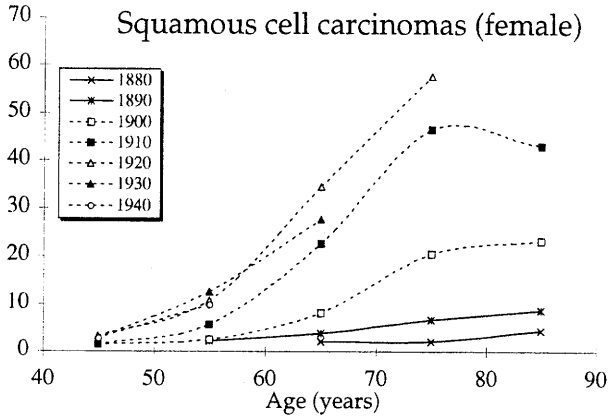
About 75% of all lung cancer cases of specified histology were squamous cell carcinomas for the earliest male cohort studied (birth year cohort – 1880s) but this fraction has decreased considerably in ensuing birth decade cohorts. Among females, who as a group showed an increase in smoking prevalence since the birthyear cohort of the 1910s, squamous cell carcinoma was never greater than 50% of all reported lung tumors and this fraction declined from a maximum in the birth cohort of the 1910s to the most recent cohort analyzed. Since only 2-3% of women born in the 1880s smoked (Harris, 1983), most lung cancer cases in this female cohort would have been expected to have occurred among the nonsmoking fraction. From these data and others it has been established lung cancer cases among nonsmokers are predominantly adenocarcinomas. (Wynder and Berg, 1967; Cooper et al, 1968; Vincent et al, 1965; Ernster, 1994) While the literature clearly defines the histopathologic spectrum of lung cancers, the position of tumors recorded with regard to position in the lung is not as well defined. Tumors of "mixed" histopathology are noted but we have not been able to ascertain, for instance, the fraction of adenocarcinomas or tumors of mixed histopathology that arise in the tracheal bronchial epithelium as opposed to the peripheral bronchiolar epithelium. Assuming that

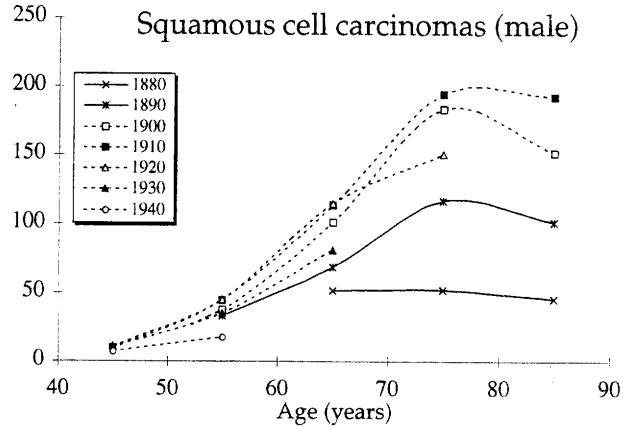
Figure 34. Age-specific lung cancer incidence rates organized by birth decade cohorts and histopathologic form of cancer

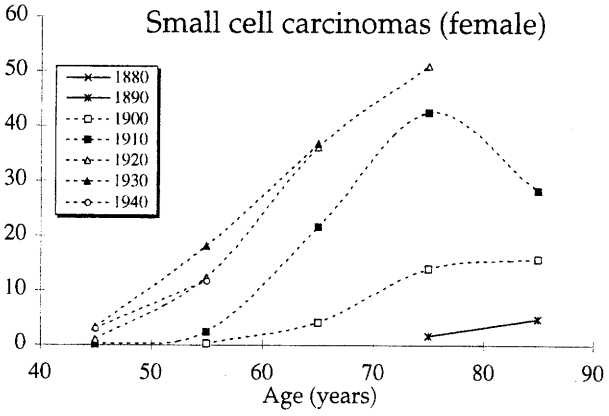
(Connecticut rates, Thun et al, 1997)

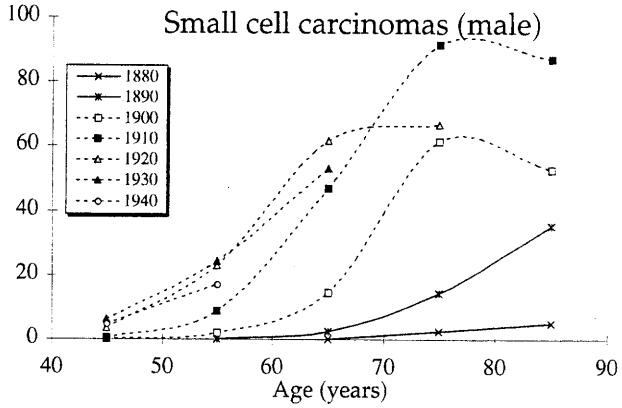












Incidence per 100,000

adenocarcinomas are all in the peripheral bronchiolar epithelium appears to be unjustified. Risk parameters for these two epithelial cell populations will be calculated using a range of estimates of the fraction of lethal lung tumors arising in the two lung regions.

3.2.9 Smoking Cessation - Lung Cancer Incidence in Former Smokers

Peto et al (2000) have collected and organized incidence data for lung cancer as a function of time after smoking cessation (Section 4.2.9). The data suggest that while smoking cessation reduced the lifetime risk of developing lung cancer in former smokers, incidence rates did not decrease to levels observed for nonsmokers. The lifetime risk for lung cancer in a former smoker is thus revealed to be not only a function of the time since smoking was ceased, but also a function of the duration that individuals smoked. Analysis of Peto et al's data by the three-stage carcinogenesis model can test the hypothesis that smoking cessation returns the individual to the lifetime risk and physiological parameters of nonsmokers soon after smoking cessation and that residual elevated risk has been created by establishment of preneoplastic colonies during the period of cigarette use.

3.3 KNOWN PHYSIOLOGICAL PARAMETERS

3.3.1 Number of cells at risk

3.3.1.1 Number of cells at risk – Growth of child

The volume of an organ is assumed to increase proportionally to the mass of an average individual. Since the colon is approximately a cylindrical tube and the lung is approximately a series of tubes, the number of colonic and bronchial epithelial cells is thereby proportional to body mass to the two-thirds power.

Figures 35a and 35b below show the masses of average males and females respectively as a function of age. For both males and females, body mass increases exponentially from age 1.5 years to 14.5 years in females and 16.5 years in males. A higher constant rate is obtained for growth between birth and age 1.5 years.

From Figures 35a and 35b, the growth rates of males and females can be estimated from the slope of the $\log_2$ of the mass of average individuals for the age intervals 0-1.5 and 1.5-14.5 for females and 1.5 to 16.5 for males. These estimated growth rates for mass were then multiplied by 2/3 to obtain estimates for the growth rates of colonic and lung epithelial cells, representing these organs as tubular.

	Ages	Growth rate (mass)	Growth rate (colonic, lung cells)
Males			
	0-1.5	1.23	0.82
	1.5-16.5	0.159	0.106
Females			
	0-1.5	1.17	0.78
	1.5-16.5	0.167	0.111

The number of colon and lung epithelial cells as a function of age, N_a, can therefore be written as a function relative to the number of colonic epithelial cells in an adult, N_{max}. For males, the number of cells as a function of age is:

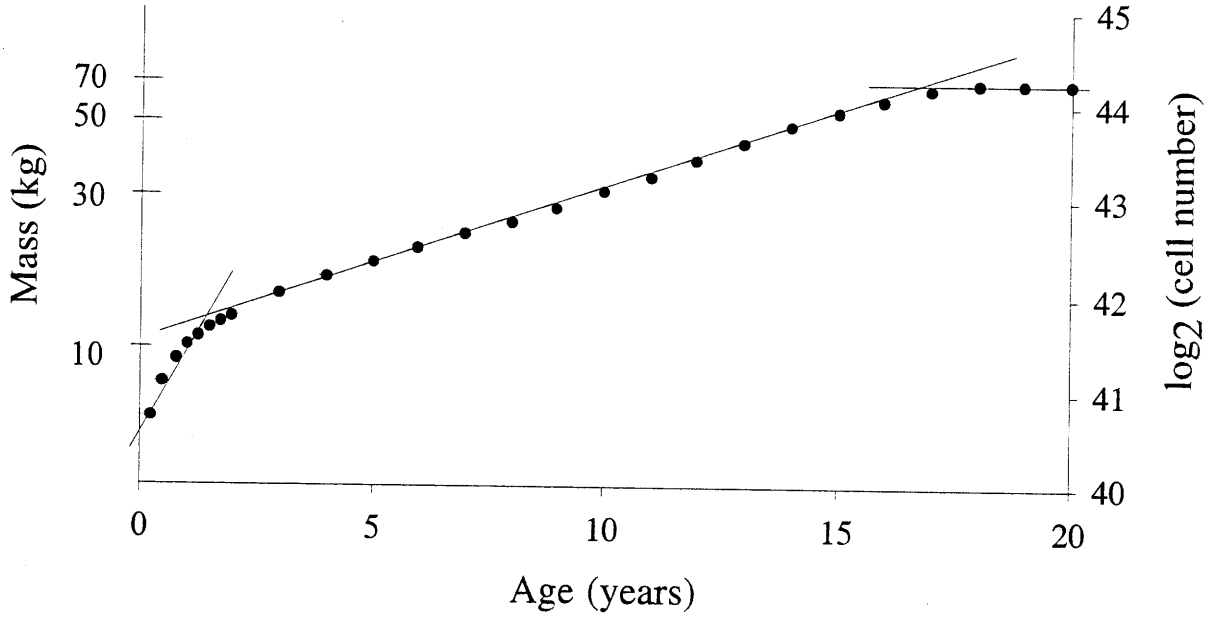
(Eq. 5)

$$N_{a,\,males} = \begin{cases} N_{max} = \text{cells in adult organ} & a > 16.5 \\ N_{max} \div 2^{0.106(16.5-a)} & 1.5 < a \leq 16.5 \\ N_{max} \div 2^{0.106(15) + 0.82(1.5 - a)} & 0 \leq a \leq 1.5 \end{cases}$$

Fig. 35: Mass of males (a) and females (b) as a function of age

(Hamill et al, 1979)

a)


Mass (kg)
70
50
30
10
0
5
10
15
20
Age (years)
45
44
43
42
41
40
log2 (cell number)

b)


Mass (kg)
70
50
30
10
0
5
10
15
20
Age (years)
45
44
43
42
41
40
log2 (cell number)

The number of cells in a female follows by similar reasoning.

One should also account for the fact that the weight of an average female is about 80% that of a male at age 18 (Hamill et al, 1979).

3.3.1.2 Number of Cells at Risk – Colon

The number of cells in an adult colon can be calculated as follows. On average, the cross-sectional count of the number of cells is 262 along the 'V-shaped' crypt. The number of cells along the circumference of the top of the crypt is approximately 40-60 cells. Assuming a crypt has a cone-like shape, then the number of cells per crypt is:

$$\sim 50 \cdot \frac{262}{2} = \frac{6550 \text{ cells}}{\text{crypt}}$$

where the size of a cone is about half of that of a cylindrical tube. Additionally, the number of crypts in any direction is about 100 crypts per cm. leading to:

$$\frac{\text{crypts}}{\text{area}} = \frac{100 \text{ crypts}}{\text{cm}} \cdot \frac{100 \text{ crypts}}{\text{cm}} = \frac{10^4 \text{ crypts}}{\text{cm}^2}$$

The colon can be approximated to be a cylinder of 130 cm in length and 10 cm in circumference leading to a surface area of 1300 cm^2. Therefore the total number of cells in the colon is:

$$\frac{\text{cells}}{\text{crypt}} \text{ x } \frac{\text{crypts}}{\text{area}} \text{ x surface area } = 6550 \cdot 10^4 \cdot 1300 \approx 8.5 \text{ x } 10^{10} \text{ cells}$$

Correcting for the different sizes of males and females leads to an estimate of $9.1 \text{ x } 10^{10}$ colonic cells in an adult male colon and $7.9 \text{ x } 10^{10}$ in an adult female.

3.3.1.3 Number of cells at risk – Lung

The number of total epithelial cells down to the sixth bronchial bifurcation estimated from *post mortem* dissections is about 2.4 x 10^8 cells. (Dr. Xiao-Cheng Li-Sucholeiki, unpublished) Most squamous cell carcinomas and small cell carcinomas are reported to arise in this region (Berg, 1970; Walter and Pryce, 1955a, 1955b). By calculations from histological preparations and the anatomy of the bronchial-alveolar region of the lung some 2.8 x 10^{10} epithelial cells line the bronchial tree from the sixth to the approximately 24^{th} bifurcation for the most extensive "trees" before the alveolar ducts and alveoli (Kuhn, 1995; Prodi and Mularoni, 1984). These numbers are employed in the models of carcinogenesis in the tracheal bronchial and bronchiolar regions of the lung respectively. The number of epithelial cells in the alveolar ducts and alveoli approaches 8 x 10^{11} but the low number of reports of tumors of alveolar origin indicates that these cells are at negligible risk of giving rise to tumors (Zheng et al, 1994).

3.3.2 Cell Kinetic Rates

Expression of mutation rates per cell division further requires knowledge of *in vivo* cell kinetic parameters.

3.3.2.1 Cell Kinetics - Colon

To count the number of mitotic events and cells undergoing apoptosis in normal tissue, precancerous lesions, and carcinomas from a long series of observations from many patients, slides can be prepared in either of two ways. For observation of mitotic figures, tissues are stained with eosin and hematoxylin. Mitotic figures can then be observed with a trained eye under a high-powered microscope. Apoptotic figures are observed after *in situ* labeling, a process

which identifies DNA with nuclear fragmentation, a feature characteristic of early apoptosis. First, microscope slides containing slices of paraffin-embedded tissue are heated to remove the surrounding wax. The tissue slices are then treated with Proteinase K to digest cellular proteins. Biotin-labeled deoxycytosine and deoxyadenosine triphosphates are combined with unlabelled deoxythymidine and deoxyguanosine triphosphates and added to the slides along with polymerase I Klenow fragment. The polymerase uses the dNTPs to fill in gaps in the DNA "ladder" that is a hallmark of apoptosis. After a short incubation period in a buffer (containing Tris-HCl, magnesium chloride, bovine serum albumin and mercaptoethanol), hydrogen peroxide and methanol are added to block endogenous peroxidase. Horseradish peroxidase conjugated to streptavidin is then applied to the slides and binds to incorporated biotin-labeled dNTPs. A chromogen, diaminobenzidine (DAB), is then applied to turn the horseradish peroxidase a color (brown) in order to visualize the conjugated and incorporated nucleotides. Under a high-powered microscope, apoptotic bodies can simply be counted by observing the morphologically small, brown pigmented cells.

Apoptotic and mitotic cell counts were done by Dr. E.E. Furth and P. Belair of the University of Pennsylvania Medicine School. The observations, expressed as events per 100 cells observed in the crypts, adenomas or carcinomas, are summarized in Figure 36. Apoptotic and mitotic counts allow estimation of division and death rates by the following transform:

(Eq . 6)

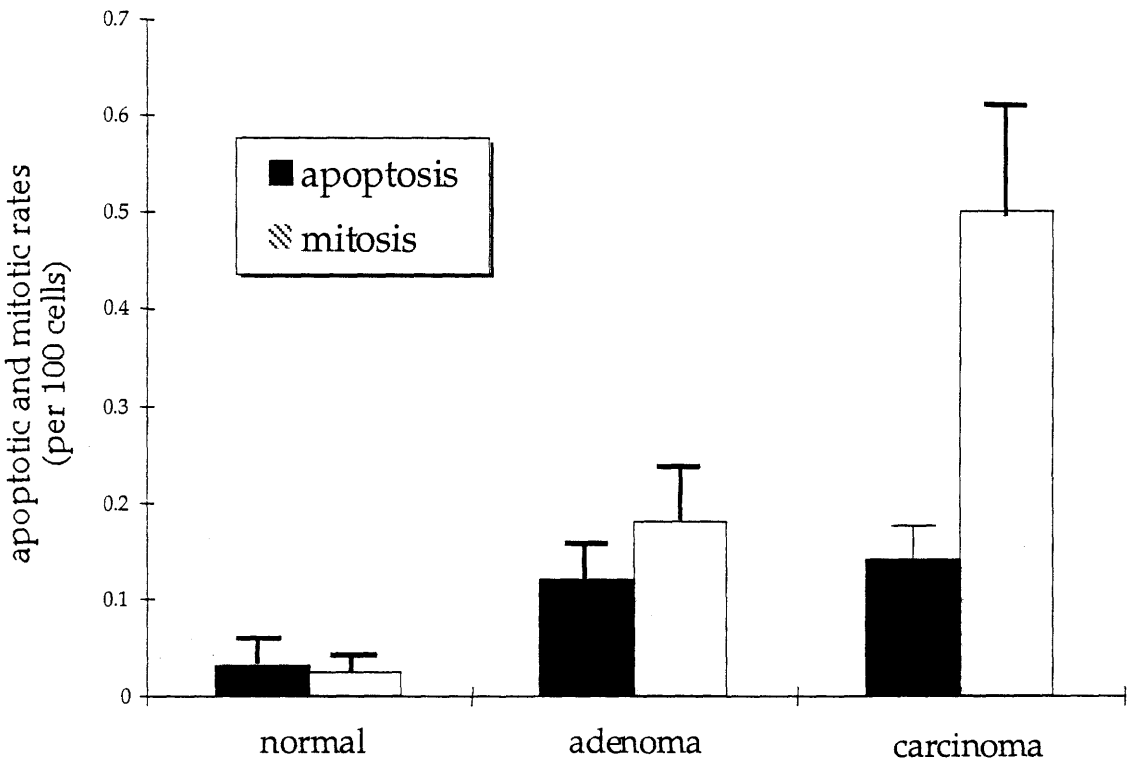
rate = (mitotic or apoptotic count) / (mitotic or apoptotic time expressed in years) x

2 if normal tissue

$$\text{i.e., normal } \alpha = \frac{.025 \text{ mito.}}{100} \cdot \frac{1}{1.5 \text{ hr.}} \cdot \frac{24 \text{ hr.}}{1 \text{ day}} \cdot \frac{365 \text{ days}}{1 \text{ yr.}} \cdot 2 = 2.92 \frac{\text{mito.}}{\text{yr.}}$$

Fig. 36: Apoptotic and mitotic cell counts of colonic tissue

Colonic Neoplastic Progression



Therefore,

	Division	Death
Normal tissue (τ):	2.92	2.92
Adenoma (α, β):	8.92	~8.92
Carcinoma (α_c, β_c):	29.2	~8.92

Mitotic times were estimated from Wright et al (1973) who reported that the duration of mitosis for the small intestine is 1 – 1.5 hours *in vivo*. Likewise, Weinstein et al (1973) found that the mitotic time for the human jejunum is 1.4 to 2.2 hours *in vivo*. (No data were found for normal colon). The lengths of mitosis and apoptosis may additionally differ among normal transition cells, adenoma and carcinoma cells. Treatment herein assumes they are in fact the same for these three cases. In normal tissue, only half of the cells can undergo division, since one half are non-dividing terminal cells. The turnover rate of normal colon epithelial cells was therefore additionally adjusted by a two-fold factor.

Adenomas have higher death and division rates, both about 9 per year. Adenomas grow slowly so the precision of such observations is not sufficient to detect a small difference between division and death rates. Small carcinomas show an identical death rate to adenomas, but the division rate is increased to about 30 per year. A difference between division and death rates of about 20 yr^{-1} implies a doubling time of about 18 days, a rapidly growing tumor.

3.3.2.2 Cell Kinetics – Lung

Several groups have reported mitotic and apoptotic indices in normal bronchial epithelial tissue, in presumptive preneoplastic colonies identified as "dysplastic or carcinomas *in situ*," and in macroscopic lung tumors.

For normal bronchial epithelium, Dr. Elena Gostjeva (Kiev Biotechnology Institute, personal communication) observed a mitotic index of about approximately 0.05% in cells sampled from the tracheal bronchial epithelium. This value leads to an estimate of 5.7 cell divisions per year among stem or transitional cells assuming a duration of mitosis of 1.5 hours in the lung. 5.7 represents the maintenance turnover rate, deaths and divisions per year, designated as "τ" in the equations describing mutation initiation rates per cell year.

Reports by Tormanen et al (1999) and Cemerikic-Martinovic et al (1998) provide apoptotic indices of dysplastic lesions, carcinomas *in situ*, and carcinomas of the lung. These data coupled with assumptions about the length of observable mitoses and apoptoses (Staunton, 1995) leads to an estimate of the rate of cell divisions and deaths of approximately 13 events per cell year for potentially preneoplastic lesions. This serves as the approximate value of both "α" and "β", respectively the division and death rates in the equations describing survival probabilities and growth rates in preneoplastic lesions. Apoptotic indices for early Grade I tumors lead to the estimate of a death rate of 28 and a division rate of 42.2 per cell year. These values are used as the estimates of the rate of cell division "α_c", the division rate, and "β_c", the death rate, in lung tumors. From these, the estimated fractions of newly initiated cells which survive to form preneoplastic colonies and promoted cells which survive to form tumors, respectively $[(\alpha - \beta) / \alpha]$ and $[(\alpha_c - \beta_c) / \alpha_c]$, are derived (Moolgavkar, 1990b).

3.4 MATHEMATICAL DEFINITIONS

3.4.1 Primary (Lifetime) Risk Factors vs. Secondary (Accelerating) Risk Factors

Here the term "primary risk" requires careful definition. Supposing that there are persons who by virtue of their genetic inheritance and environmental experience are at risk of cancer,

a subpopulation at primary risk for cancer would thereby exist. It is possible that the entire population has the same genetic risk but not the same environmental experience. Conversely, it is possible that a common environmental experience is shared by all persons, but only a fraction carry an inherited risk factor. The key postulate is that persons who do not inherit and experience these primary risk factors cannot develop cancer in a full lifetime of up to, say, 125 years. Primary genetic and environmental risk factors for sporadic colon cancer have not yet been identified and are, therefore, hypothetical. (The primary genetic risk factors for two forms of familial colon cancer, FAPC and Lynch syndrome (HNPCC), are an inactive allele of the APC gene or of a mismatch repair gene respectively) (Kinzler et al, 1991; Leach et al, 1993). Furthermore it is not clear whether cigarette smoke acts as a primary risk factor, essential in the development of lung cancer, or rather accelerates the onset of lung cancer among individuals already independently at primary risk for cancer, or possibly both.

Within each subpopulation at primary risk, variations in mutation rates and cell kinetic rates are to be expected. When an inherited condition or environmental experience lowers the expected age of death relative to all persons at primary risk, it describes a secondary risk factor, accelerating the process by which an individual can die of cancer. For instance, persons with mutation rates only twofold higher than average would be expected to develop cancers much earlier in life than persons with average mutation rates within the subpopulation sharing the same primary risk factors, as they would accumulate all of the necessary initiation and/or promotion events at a faster rate. Inherited or environmental factors affecting mutation rates or precancerous growth rate would, by this definition, be secondary risk factors.

3.4.2 Subpopulations at risk

The data of Figures 25 through 28 comprise all recorded deaths from intestinal cancers and lung cancers. For the case of intestinal cancer, when survival and underreporting are accounted, Figures 31 and 32, it is clear by inspection that these functions reach a maximum in old age. This repeated observation is consistent with expectation for a population in which only some fraction is at lifetime risk of cancer. While recognizing that other explanations for such a maximum may be devised, the analysis herein is built on the validity of the subpopulation at risk assumption and the certain knowledge that human populations display a high degree of genetic heterogeneity.

These data do not, however, separate deaths in families with familial adenomatous polyposis coli (FAPC) from deaths in families with hereditary nonpolyposis colon cancer (HNPCC or Lynch syndrome) or from deaths by "sporadic" colon cancer. "Sporadic" cancers themselves are undifferentiated with regard to the possibility that there are independent pathways of genetic changes leading to several different kinds of "sporadic" cancer. Also by example, the lung cancer mortality data of Figures 27 and 28 do not separate by squamous cell lung carcinomas, small cell lung carcinomas, and adenocarcinomas.

There could be multiple pathways to cancer in any particular organ. The potential for and rate of transit of these pathways would be determined by unknown but ascertainable alleles of tumor suppressor genes and genes which effect the rates of genetic changes and cell kinetic rates in normal tissues and preneoplastic colonies. These alleles would be distributed throughout the entire population.

There are cancers of organs for which such a treatment assuming multiple pathways is obviously required. Figure 37 shows OBS(h,t) for death by testicular cancer in which two

populations are clearly evident, one with all deaths occurring between ages 15 and 40 and a second group in which deaths begin to be observed after age 50. Mortality data lump the deaths from multiple independent pathways together perforce.

Even if many possible pathways exist to mortal cancer inherent in a particular individual, death can be caused by only one. Thus the number of cancer deaths must be the sum of the deaths caused by each of the potentially multiple pathways:

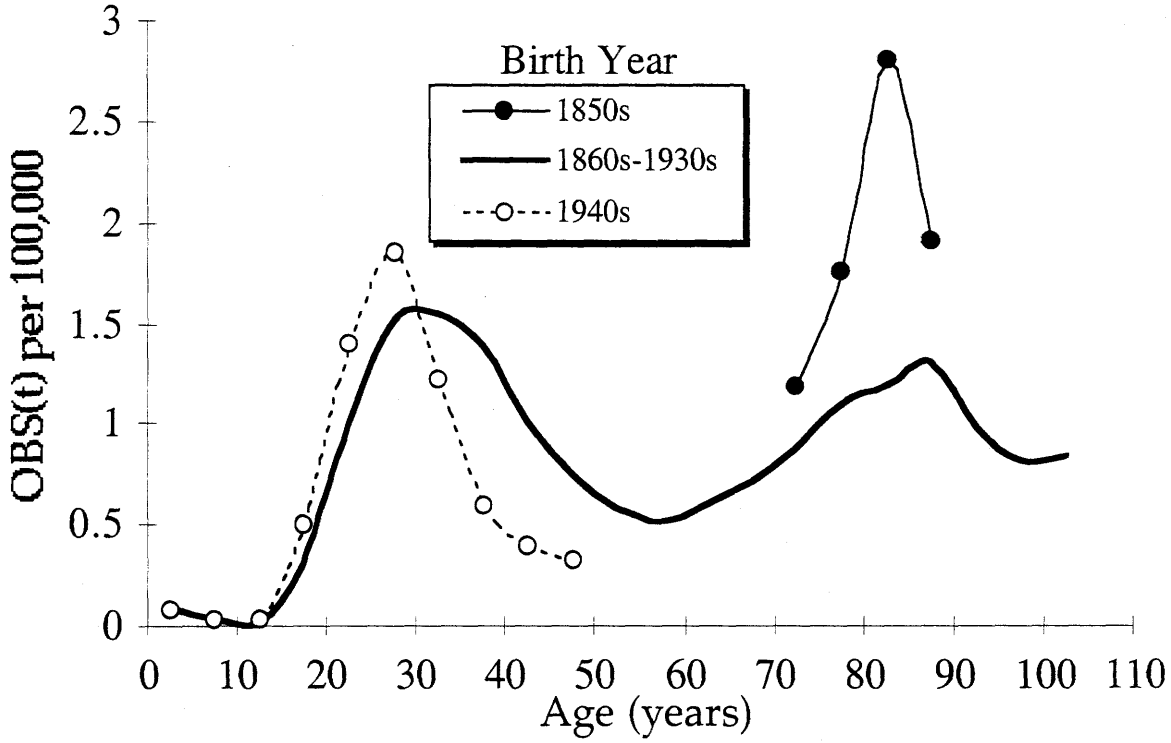
(Eq. 7) $OBS(h,t) = OBS_1(h,t) + OBS_2(h,t) + OBS_3(h,t) + ...$

In the case of colon cancer, mortality from FAPC and HNPCC families is numerically small and occurs earlier in life than the "sporadic" form(s) of the disease. For the time being their real but numerically small contribution to total colon cancer mortality is neglected. Some 80% of colorectal adenomas in FAPC individuals have been found to lack an operative APC allele. It appears, therefore, that "sporadic" colon cancers have a single, common initiation pathway, loss of the two inherited operative alleles of the tumor suppressor gene APC (Powell, 1992). It is also tempting to assume that the genetic change(s) needed in the promotion of a "sporadic" precancerous cell to a carcinoma cell would be the same for all individuals, but this assumption is without any evidentiary support and unnecessary for the analyses attempted below.

These points being noted, colon mortality data can be initially modeled *as if* there were one and only one pathway to colon cancer. If there are multiple pathways to "sporadic" cancer, the derived parameters such as the number of mutations required in initiation and promotion, their rates and the growth rate of adenomas are, perforce, a weighted average among the multiple pathways.

Fig. 37: Testicular cancer age- and birthyear-specific mortality, EAM.

Evidence of independent populations at risk for cancer, or independent pathways of carcinogenesis: one between the ages 15 and 40, and the other later in life after age 60


Birth Year
1850s
1860s-1930s
1940s
OBS(t) per 100,000
3
2.5
2
1.5
1
0.5
0
0
10
20
30
40
50
60
70
80
90
100
110
Age (years)

Independent pathways to cancer could also represent independent causes. For example, in the case of lung cancer, deaths could be subdivided into at least two groups: deaths caused by smoking cigarettes, and deaths caused by all other potential risk factors. Perforce the possibility that there exist two separate, but not necessarily mutually exclusive subpopulations at risk for lung cancer must be considered.

3.4.3 Definition of Primary Risk Fraction

The age-specific cancer mortality for any birth year cohort, OBS(h,t), can be expressed as a function of the primary and secondary risk factors. The fraction of the population at primary risk within a birth year cohort, F(h,t), where "h" is the historical birth year and "t" is age, is defined by the interaction of inherited and environmental primary risk factors:

(Eq. 8) $$F(h,t) = F(h,t)_{genetic} \times F(h,t)_{environmental}$$

This is an important abstraction: the fraction of the cohort that would die of cancer if there were no other causes of death. [F(h,t) is <u>not</u> the fraction of the cohort observed to die of cancer which is much smaller].

Assuming that there is little historical variation in the fraction of the population inheriting primary genetic risk factors of the nearly 100 years for which data was acquired, $F(h,t)_{genetic}$ = G is a constant. Thus, any real change in F(h,t) with "h" would be ascribed to historical changes in the environmental primary risk factor, $F(h,t)_{environmental}$.

Since historical changes in environmental factors would rarely reach all of the population simultaneously, primary environmental factors can vary significantly within the lifetimes of some birth year cohorts, e.g. the cohorts for whom manufactured cigarettes were not available until middle age. At this stage of model development, however, F(h,t), will be modeled as invariant within a birthyear cohort in the case of colon cancer, such that:

(Eq. 9) $F(h,t) \approx \mathbf{F_h} = G \times E_h$

The idea of a fraction of the population with both inherited and environmental risk factors for cancer is logically straightforward. That fraction is represented as $G\ E_h$. But it follows that there also exist three other distinct subpopulations: those that have neither risk factor, $(1 - G)\ (1-E_h)$, those that have the environmental but not the inherited risk, $(1 - G)\ E_h$, and those that have the inherited but not the environmental risk, $G\ (1-E_h)$. This point is illustrated in a Venn diagram (Figure 38). Each of these subfractions would have potentially different age-specific death rates.

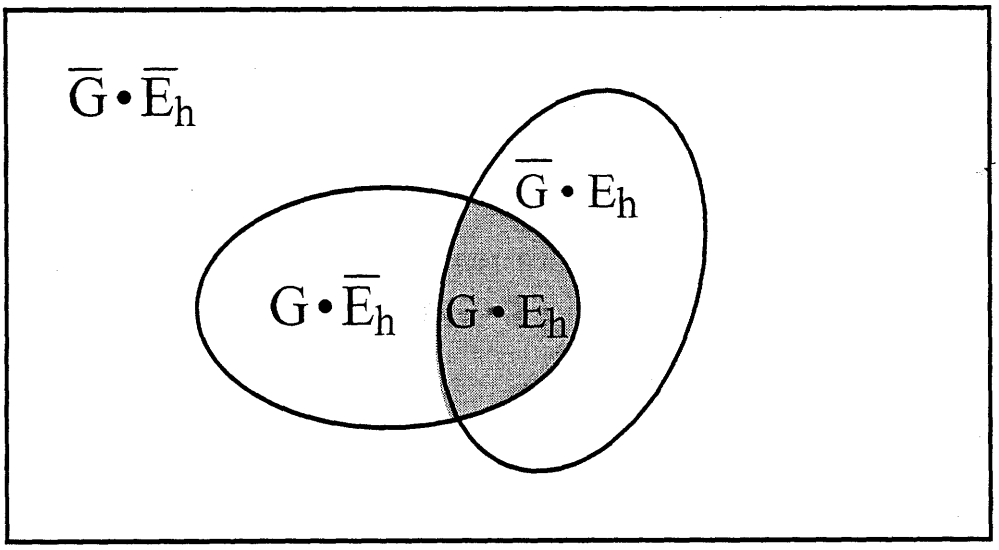
3.4.4 Definition of Causes of Death Given Inheritance and/or Exposure to a Primary Risk Factor

The total number of deaths within an age interval in any historical year is the sum of the number of deaths from all possible causes. Thus, the total recorded mortality rate, TOT(h,t), is the sum of the rate of deaths by the cancer of interest, OBS(h,t), the rate of deaths from connected causes sharing the same primary genetic and/or environmental risk factors as that cancer, CON(h,t), and the rate of deaths from causes independent of the primary genetic or environmental causes of that cancer, IND(h,t). Algebraically:

(Eq. 10) $TOT(h,t) = OBS(h,t) + CON(h,t) + IND(h,t)$

Fig. 38: Venn diagram of the population at primary risk as defined by genetic and environmental risk factors.

Representation of population at risk, F(h,t), as the intersection of the population at genetic primary risk (G) and the population at environmental primary risk (E_h)



$G = F(h,t)_{genetic}$
$E_h = F(h,t)_{environmental}$

$\overline{G} = (1 - G)$ = not in $F(h,t)_{genetic}$
$\overline{E}_h = (1 - E_h)$ = not in $F(h,t)_{environmental}$

The historical record defines estimates of TOT(h,t) and OBS(h,t) whereas the values of CON(h,t) and IND(h,t) are unknown.

For each of these categories of mortality there is a related age-dependent hazard function. These probabilities are abstract, age-dependent functions:

$P_{OBS}(h,t)$ = probability that a person born in year 'h' having both the primary genetic and environmental risk for the cancer of interest would die of that cancer at age 't' given no treatment and no competing forms of death. (population at risk: $G \bullet E_h$, Figure 38). **Unreported** colon cancer deaths are included in this category.

$P_{CON}(h,t)$ = probability that a person born in year 'h' having either primary genetic and/or environmental risk of the cancer of interest would die by any form of death connected to either or both of these risks other than the cancer of observation at age 't', given no treatment and no competing independent forms of death. (populations at risk: $G \bullet E_h$, $G \bullet \overline{E_h}$, and $\overline{G} \bullet E_h$, Figure 38)

$P_{IND}(h,t)$ = probability that a person born in year 'h' having neither a primary genetic nor environmental risk of colon cancer would die of any other cause at age 't'. (all populations at equal risk: Figure 38)

$OBS^*(h,t) = OBS(h,t) \div [R(h,t)\ (1 - S(h,t))]$ therefore represents the **observed** recorded cancer mortality rate for individuals born in year 'h' who were still alive at age 't' in an abstract world without therapeutic treatment, while **$P_{OBS}(h,t)$** is the **expected** cancer mortality rate for an

individual belonging to the group $\mathbf{F_h}$ who is still alive at age 't', given no medical intervention. OBS*(h,t) will suffer from errors in reporting and diagnosis not accounted by our use of R(h,t), but $\mathbf{P_{OBS}(h,t)}$, the derived mortality probability for persons in the subpopulation $G \bullet E_h$, represents all deaths by the cancer of interest whether they are diagnosed and/or reported accurately or not in an abstract world where S(h,t) = 0.

The term for the actual probability of death from colon cancer for a person at risk of birth year cohort h and age t would be the probability of dying of colon cancer in the absence of treatment, $\mathbf{P_{OBS}(h,t)}$, multiplied by the probability that treatment has not been successful (1 - S(h,t)). Similar arguments can be introduced for the connected and independent forms of mortality so that the probability of not surviving a 'connected' disease would be $(1 - S_{CON}(h,t))$ and of an 'independent' form of death, $(1 - S_{IND}(h,t))$.

3.5 ALGEBRAIC EXPRESSIONS

3.5.1 Probability of Still Being Alive at Age 't'

We assume that a person at primary risk of the cancer of interest can die of that cancer, a 'connected' disease caused by either the inherited and/or environmental risk factors, or a cause independent of the primary risk factors for that cancer. This permits the writing of the explicit statement for the probability, $P_{NOT}(h,t)$, that a person within the risk group $\mathbf{F_h}$ has not yet died from any cause at any age between birth and age 't'.

(Eq. 11)

$$P_{NOT}(h,t) = e^{-\int_0^t [\mathbf{P_{OBS}(h,t)}\,(1-S(h,t)) + P_{CON}(h,t)\,(1-S_{CON}(h,t)) + P_{IND}(h,t)\,(1-S_{IND}(h,t))]\,dt}$$

This expression is important because in considering the probability of death by cancer at age 't', one required physical condition is that the individual not be already dead. In the terminology of probability and analysis, we are writing equations for the conditional probability of colon cancer given the fact that the individual is not dead. In considering the diminution of the subpopulation $\mathbf{F_h} = G \bullet E_h$, we are using a probability model of sampling without replacement.

3.5.2 Observed Mortality Rate at Age 't', OBS(h,t)

3.5.2.1 Fraction at Primary Risk

As described in Section 3.4.3, a cohort is comprised of 4 subpopulations, the fraction at risk for the cancer of interest, the fraction at risk of dying of the forms of death sharing primary genetic, but not environmental, risk factors with that cancer, the fraction at risk of dying of the forms of death sharing primary environmental, but not genetic, risk factors with that cancer, and the fraction at risk of dying of only forms of death independent of all risk factors of that cancer. The forms of death by which each subpopulation of Figure 38 is at risk are:

GE_h	$\overline{G}E_h$	$G\overline{E}_h$	$\overline{G}\overline{E}_h$
OBS			
CON	CON1*	CON2*	
IND	IND	IND	IND

* All forms of death in either CON1 or CON2 are included in the entire set of connected forms of death, CON.

Accounting for the effects of survival, S(h,t), underreporting error, R(h,t), and the four distinct populations introduced above, the complete equation for the observed mortality of the desired disease, OBS(h,t), may be written as follows

(Eq. 12)

$$OBS(h,t) = \frac{B_h \cdot (G \cdot E_h) \cdot (1 - S(h,t)) \cdot R(h,t) \cdot \mathbf{P_{OBS}(h,t)} \cdot P_{NOT}(h,t)}{B_h \cdot [G \cdot E_h \cdot P_{NOT}(h,t) + (1 - G) \cdot E_h \cdot e^{-\int_0^t [P_{CON1}(h,t)(1 - S_{CON1}(h,t)) + P_{IND}(h,t)(1 - S_{IND}(h,t))]\,dt} + G \cdot (1 - E_h) \cdot e^{-\int_0^t [P_{CON2}(h,t)(1 - S_{CON2}(h,t)) + P_{IND}(h,t)(1 - S_{IND}(h,t))]\,dt} + (1 - G) \cdot (1 - E_h) \cdot e^{-\int_0^t P_{IND}(h,t)(1 - S_{IND}(h,t))\,dt}]}$$

OBS(h,t) equals the number of persons within a birth cohort 'h' who are recorded as dying of the cancer of interest at age 't' divided by the number of all persons in the cohort still alive at that age.

The **numerator**, the number of deaths from the cancer of interest at age 't', is the product of the number of persons in the cohort at birth, B_h, the fraction at primary risk, (G x E_h), the fraction of individuals who develop that cancer and do not survive, (1 - S(h,t)), the estimated fraction of these cancer deaths accurately recorded, R(h,t), the fraction expected to die of that cancer in the absence of treatment, $\mathbf{P_{OBS}(h,t)}$, and the fraction of (G x E_h) not already dead from any cause, $P_{NOT}(h,t)$.

The **denominator**, the number of persons still alive at age 't', is the product of the number of persons in the cohort at birth, B_h, and the sum of the fractions of all subpopulations still alive, adjusted by each subpopulation's specific probability of not being already dead.

Since the terms for number of persons born to a cohort, B_h, and the term accounting for survival from causes unrelated to the cancer risk factors, $e^{-\int P_{IND}(h,t)\,(1 - S_{IND}(h,t))\,dt}$, are present as factors in the numerator and all terms of the denominator, they cancel out. The algebraic elimination of the term for the probability of independent forms of death is extremely important since there is no satisfactory way of determining its value from public mortality records. By next dividing the numerator and denominator by $e^{-\int [P_{OBS}(h,t)\,(1 - S(h,t) + P_{CON}(h,t)\,(1 - S_{CON}(h,t))]\,dt}$, we convert this equation into a more manageable form.

(Eq. 13)

$$OBS(h,t) =$$

$$\frac{(G \cdot E_h) \cdot (1 - S(h,t)) \cdot R(h,t) \cdot \mathbf{P_{OBS}(h,t)}}{\left[G \cdot E_h \cdot e^{-\int_0^t [\mathbf{P_{OBS}(h,t)}\,(1 - S(h,t)) + P_{CON}(h,t)\,(1 - S_{CON}(h,t)) - P_{CON1}(h,t)\,(1 - S_{CON1}(h,t))]\,dt} + (1 - G) \cdot E_h \cdot e^{-\int_0^t [\mathbf{P_{OBS}(h,t)}\,(1 - S(h,t)) + P_{CON}(h,t)\,(1 - S_{CON}(h,t)) - P_{CON2}(h,t)\,(1 - S_{CON2}(h,t))]\,dt} + G \cdot (1 - E_h) \cdot e^{-\int_0^t [\mathbf{P_{OBS}(h,t)}\,(1 - S(h,t)) + P_{CON}(h,t)\,(1 - S_{CON}(h,t))]\,dt} + (1 - G) \cdot (1 - E_h) \cdot e^{-\int_0^t [\mathbf{P_{OBS}(h,t)}\,(1 - S(h,t))]\,dt} \right]}$$

3.5.2.2 Accounting for Deaths by Causes Connected to Primary Risk Factors

"Connected" diseases are unknown and therefore there is no way of describing what $P_{CON}(h,t)$, $P_{CON1}(h,t)$, $P_{CON2}(h,t)$, $S_{CON}(h,t)$, $S_{CON1}(h,t)$ or $S_{CON2}(h,t)$ might be. To move beyond this clear absence of data requires an algebraic approximation.

We define the term f(h,t) as the ratio of deaths by the cancer of interest to all deaths actually caused by either the inherited or environmental risk factors for that cancer of interest. This fraction, f(h,t), is assumed to be constant for all ages within a birth year cohort but may vary among birth year cohorts, such that $f(h,t) \approx \mathbf{f_h}$. Assuming an age-independent ratio is not necessarily grossly improper, as CON1 and CON2 may include other forms of cancer which might therefore have an age dependence and survival probability similar to that of the cancer of interest. In balance: the relative age dependence of cancer mortality is not greatly different from the major causes of human mortality (Figure 1), vascular disease (Figure 2) and any other cancer (See http://cehs4.mit.edu).

(Eq. 14)

$$(1 - G \cdot E_h) \cdot e^{\frac{1}{f(h,t)} \int_0^t \mathbf{P_{OBS}(h,t)} (1 - S(h,t))\, dt}$$

$$\approx$$

$$\left[\begin{array}{l} (1 - G) \cdot E_h \cdot e^{\int_0^t [\mathbf{P_{OBS}(h,t)} (1 - S(h,t)) + P_{CON}(h,t) (1 - S_{CON}(h,t)) - P_{CON1}(h,t) (1 - S_{CON1}(h,t))]\, dt} \\ + G \cdot (1 - E_h) \cdot e^{\int_0^t [\mathbf{P_{OBS}(h,t)} (1 - S(h,t)) + P_{CON}(h,t) (1 - S_{CON}(h,t)) - P_{CON2}(h,t) (1 - S_{CON2}(h,t))]\, dt} \\ + (1 - G) \cdot (1 - E_h) \cdot e^{\int_0^t [\mathbf{P_{OBS}(h,t)} (1 - S(h,t)) + P_{CON}(h,t) (1 - S_{CON}(h,t))]\, dt} \end{array} \right]$$

Equation 14 defines the approximation using f(h,t) in the context of Equation 13, distributing the effects of differential death rates among the three populations not at risk for the OBServed cancer.

It is helpful to recall that the sum of all four subpopulation fractions is equal to one:

(Eq. 15)

$$[(1 - G) \times E_h] + [G \times (1 - E_h)] + [(1 - G) \times (1 - E_h)] = 1 - (G \times E_h) = (1 - \mathbf{F_h})$$

Combining Equations 13-15 creates the relatively simple expression:

(Eq. 16)

$$OBS(h,t) = \frac{\mathbf{F_h} \cdot (1 - S(h,t)) \cdot R(h,t) \cdot \mathbf{P_{OBS}(h,t)}}{\mathbf{F_h} + (1 - \mathbf{F_h}) \cdot e^{\frac{1}{\mathbf{f_h}} \int_0^t \mathbf{P_{OBS}(h,t)} (1 - S(h,t))\, dt}}$$

for which all values are known save for $\mathbf{F_h}$, $\mathbf{f_h}$ and $\mathbf{P_{OBS}(h,t)}$ shown in bold face. Sections 3.5.4 and 3.5.5 demonstrate how to explicitly solve for these parameters.

3.5.3 Observed lung cancer mortality rate at age t, OBS(h,t), of a mixed population of smokers and nonsmokers

Real populations consist of smokers and nonsmokers. Three groups are at potential risk for lung cancer:

(a) smokers at risk only because of smoking

(b) nonsmokers at risk because of unknown factors unrelated to smoking

(c) smokers at risk because of smoking and unknown factors unrelated to smoking

The observed mortality rate, OBS(h,t), can thus be written as a function of each of the three subpopulations (subscripts: NS - nonsmoker, and S - smoker):

(Eq. 17)

$$OBS(h,t) = \frac{\left[\begin{array}{l} F_{h,NS}\cdot(1-F_{h,S})\cdot P_{OBS,NS}(h,t)\cdot e^{-\frac{1}{f_{h,NS}}\int P_{OBS,NS}(h,t)\,dt} + \\ F_{h,S}\cdot(1-F_{h,NS})\cdot P_{OBS,S}(h,t)\cdot e^{-\frac{1}{f_{h,S}}\int P_{OBS,S}(h,t)\,dt} + \\ F_{h,NS}\cdot F_{h,S}\cdot(P_{OBS,NS}(h,t)+P_{OBS,NS}(h,t))\cdot e^{-\int(\frac{1}{f_{h,NS}}P_{OBS,NS}(h,t)+\frac{1}{f_{h,S}}P_{OBS,S}(h,t))\,dt} \end{array}\right]}{\left[\begin{array}{l} F_{h,NS}\cdot(1-F_{h,S})\cdot e^{-\frac{1}{f_{h,NS}}\int P_{OBS,NS}(h,t)\,dt} + \\ F_{h,S}\cdot(1-F_{h,NS})\cdot e^{-\frac{1}{f_{h,S}}\int P_{OBS,S}(h,t)\,dt} + \\ F_{h,NS}\cdot F_{h,S}\cdot e^{-\int(\frac{1}{f_{h,NS}}P_{OBS,NS}(h,t)+\frac{1}{f_{h,S}}P_{OBS,S}(h,t))\,dt} + \\ (1-F_{h,NS}-F_{h,S}) \end{array}\right]}$$

This "semi-hairy" expression is easily calculated using parametric values for nonsmokers derived from birth year cohorts which had not yet adopted the cigarette habit and those for populations in which the overwhelming fraction at risk was defined by smokers.

3.5.4 Explicit Terms for Primary Risk Factors, $\mathbf{F_h}$ and $\mathbf{f_h}$, for a Given Number of Initiation Mutations, 'n'

The first tactic is to introduce a simple function for $\mathbf{P_{OBS}(h,t)}$. Following the original logic of Nordling (1953) who noted that the age dependence of phenomena requiring 'n' mutations in the same cell in a cell population of constant size would rise as a function of age to the power of (n-1):

(Eq. 18)

$$\text{Nordling } \mathbf{P_{OBS}(h,t)} = K_h t^{n-1}$$

Here, K_h is a constant proportional to the product of the 'n' mutational rates and the number of cells at risk. It is, in fact, the rate of initiation, a fact useful later in deriving estimates of initiation mutation rates.

Nordling's model however does not permit for the proliferative capacity of a precancerous cell. The simplest way to incorporate the effect of promotion on cancer mortality curves is to modify the Nordling (1953) model with a time delay, Δ_h, representing the average latency time between initiation of a normal cell and the promotion of any precancerous cell into a malignant form. The modified model becomes:

(Eq. 19)

$$\text{Modified Nordling } \mathbf{P_{OBS}(h,t)} = K_h (t - \Delta_h)^{n-1} \quad (t > \Delta_h)$$

Substituting Equation 19 into Equation 16, for a given value of 'n', there are four unknown parameters: K_h, Δ_h, $\mathbf{F_h}$, and $\mathbf{f_h}$.

(Eq. 20)

$$OBS(h,t) = \frac{F_h \cdot (1 - S(h,t)) \cdot R(h,t) \cdot K_h \cdot (t - \Delta_h)^{n-1}}{F_h + (1 - F_h) \cdot e^{\frac{1}{f_h}\int_0^t K_h \cdot (t - \Delta_h)^{n-1} (1 - S(h,t))\, dt}} \qquad (t > \Delta_h)$$

To explicitly solve for any or all of these four unknown terms, four independent equations are needed.

3.5.4.1 Parameters determined by inspection: (F_h K_h) and Δ_h.

Δ_h and (F_h κ_h) are determined for each birth year cohort by inspection of the mortality data corrected for survival and under-reporting, OBS* (h,t), as illustrated in Figures 27 and 28 for lung cancer, and Figures 31 and 32 for colon cancer. To accomplish these estimations OBS*(h,t) for values up to the age t_{max} when OBS*(h,t) reaches a maximum is suitably approximated for any number of initiation mutations, n, as:

(Eq. 21) $$OBS^*(h,t) \approx F_h P_{OBS}(h,t) \approx F_h K_h (t - \Delta_h)^{n-1} \qquad (t_{max} > t > \Delta_h)$$

This formulation is essentially that first suggested by Nordling (1953) who did not have data for extreme old age and thus could not have recognized the maximum.

Its application is straightforward. For example, in the case of n = 2, the approximation of OBS*(h,t) would have a 'value' of zero up to $t = \Delta_h$ and then rise linearly with slope $(\mathbf{F_h}\ \mathbf{K_h})$. The x-intercept of the line is Δ_h. As OBS*(h,t) approaches its maximum at t_{max}, the approximation fails. Figure 39 shows the derivation of $(\mathbf{F_h}\ \mathbf{K_h})$ for colon cancer mortality rates among European-American males born in the 1870s.

Δ_h and $(\mathbf{F_h},\mathbf{K_h})$ can be similarly determined by inspection for all other values of n by simply plotting OBS*(h,t) versus t^{n-1}. For the general model with 'n' initiation mutations, Δ_h is simply the x-intercept and $(\mathbf{F_h},\mathbf{K_h})$ is the slope of the linear portion of the plot of OBS*(h,t) versus t^{n-1}.

3.5.4.2 Use of the Area Under $OBS^R(h,t)$ to Define $\mathbf{F_h}$ in Terms of $\mathbf{f_h}$.

Explicit solution of the two remaining unknowns, $\mathbf{F_h}$ and $\mathbf{f_h}$ requires two additional independent equations. The first is supplied by the integral of the equation $OBS^R(h,t)$ vs. t from t = 0 to infinity, where $OBS^R(h,t)$ is defined as the expected mortality rate if all deaths from the cancer of interest were accurately reported.

(Eq. 22) $$OBS^R(h,t) = OBS(h,t) \div R(h,t)$$

The integral of Equation 22 must equal the area observed under the curve $OBS^R(h,t)$ vs. t as illustrated in Figure 40.

Fig. 39: Estimation of parameters $(\mathbf{F_h, K_h})$ and Δ_h from linear portion of OBS*(h,t) vs. t^{n-1}

(Data illustrated: colon cancer, 1870s EAM)

x - intercept of a line drawn through the linear region of the mortality curve = Δ_h

slope of the linear region of the mortality curve ~ $(\mathbf{F_h, K_h})$

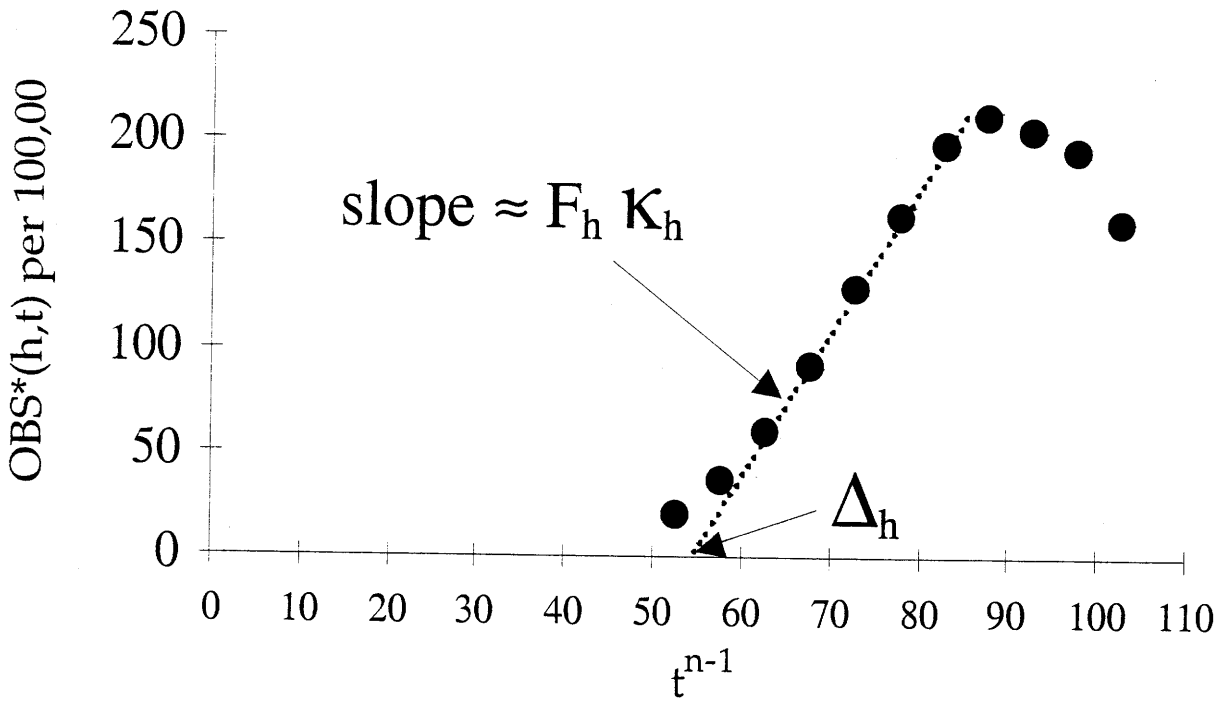
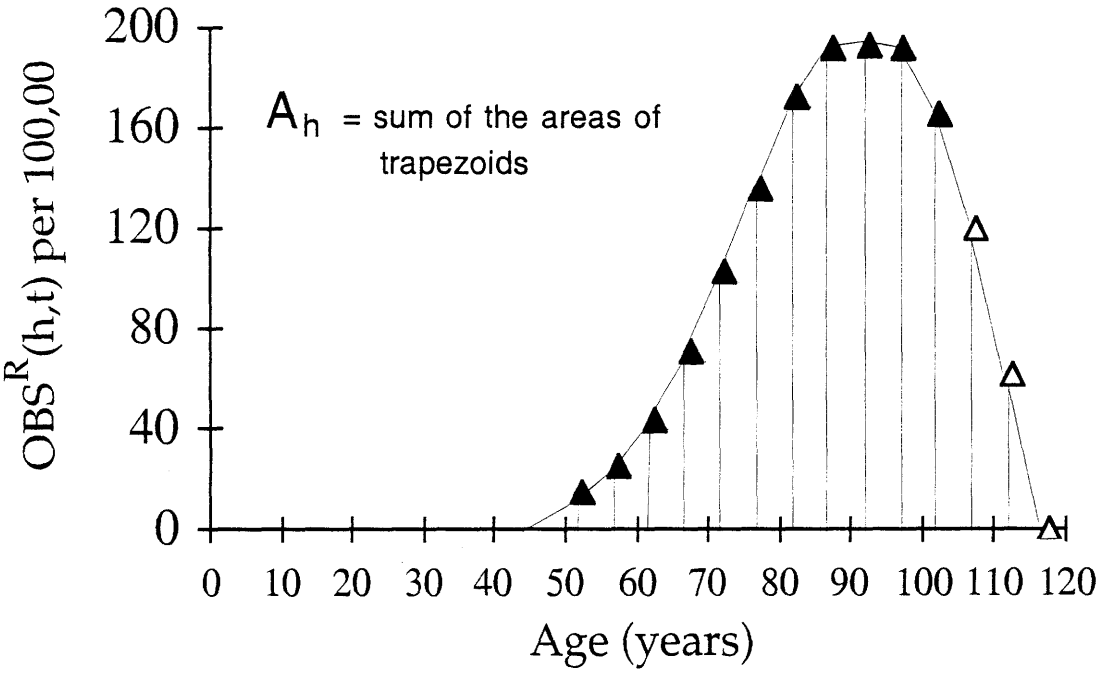

250
200
150
100
50
0
OBS*(h,t) per 100,00
slope ≈ F_h K_h
Δ_h
0
10
20
30
40
50
60
70
80
90
100
110
t^{n-1}

Fig. 40: Estimation of the integral of $OBS^{R}(h,t) = OBS(h,t) \div R(h,t)$.

Open symbols represent extrapolation of the data used for the approximation.

(Data illustrated: colon cancer, 1870s EAM)


200
160
120
80
40
0
$OBS^R(h,t)$ per 100,00
A_h = sum of the areas of trapezoids
0 10 20 30 40 50 60 70 80 90 100 110 120
Age (years)

Intuitively, this area, A_h, must be a function of the population at risk of developing the cancer of interest, $\mathbf{F_h}$, and the fraction of this population at risk which would die from that cancer, $\mathbf{f_h}$, if an individual at risk for that cancer could only die of but those forms of deaths sharing the same risk factors as the cancer of interest. This area would be independent of factors which would affect the age at which deaths are expected, such as survival rates, mutation rates and cell kinetics rates, as well as the number of events required for initiation or promotion.

The algebraic relationship between the area under $OBS^R(h,t)$ and $\mathbf{F_h}$ and $\mathbf{f_h}$ is derived as:

$$A_h = \int_0^\infty OBS^R(h,t)\,dt = \int_0^\infty \frac{\mathbf{F_h} \cdot (1 - S(h,t)) \cdot \mathbf{P_{OBS}(h,t)}}{\mathbf{F_h} + (1 - \mathbf{F_h}) \cdot e^{\frac{1}{\mathbf{f_h}} \int_0^t (1 - S(h,t)) \cdot \mathbf{P_{OBS}(h,t)}\,dt}}\,dt$$

$$= \int_0^\infty \frac{\mathbf{F_h} \cdot (1 - S(h,t)) \cdot \mathbf{P_{OBS}(h,t)} \cdot e^{-\frac{1}{\mathbf{f_h}} \int_0^t (1 - S(h,t)) \cdot \mathbf{P_{OBS}(h,t)}\,dt}}{\mathbf{F_h} \cdot e^{-\frac{1}{\mathbf{f_h}} \int_0^t (1 - S(h,t)) \cdot \mathbf{P_{OBS}(h,t)}\,dt} + (1 - \mathbf{F_h})}\,dt$$

In order to permit integration, we introduce the variable v such that:

$$v = e^{-\frac{1}{\mathbf{f_h}} \int_0^t (1 - S(h,t))\, \mathbf{P_{OBS}(h,t)}\,dt}$$

$$\frac{dv}{dt} = -\frac{1}{\mathbf{f_h}} \cdot (1 - S(h,t)) \cdot \mathbf{P_{OBS}(h,t)} \cdot e^{-\frac{1}{\mathbf{f_h}} \int_0^t (1 - S(h,t))\, \mathbf{P_{OBS}(h,t)}\,dt}$$

thus creating a simpler expression allowing solution of the definite integral:

(Eq. 23)

$$A_h = \int_0^{\infty} \frac{OBS(h,t)}{R(h,t)}\,dt = -\int_1^0 \frac{\mathbf{F_h} \cdot \mathbf{f_h}}{\mathbf{F_h} \cdot v + (1 - \mathbf{F_h})}\,dv = \int_0^1 \frac{\mathbf{F_h} \cdot \mathbf{f_h}}{\mathbf{F_h} \cdot v + (1 - \mathbf{F_h})}\,dv$$

$$= \mathbf{f_h} \cdot \ln(\mathbf{F_h} \cdot v + (1 - \mathbf{F_h}))\Big|_0^1 = \mathbf{f_h} \cdot \ln(\mathbf{F_h} + (1 - \mathbf{F_h})) - \mathbf{f_h} \cdot \ln(1 - \mathbf{F_h})$$

$$= -\mathbf{f_h} \cdot \ln(1 - \mathbf{F_h})$$

Note that by first dividing R(h,t) from OBS(h,t) avoids the need to characterize R(h,t), which is itself not explicitly integrable.

The simplicity of this result provides a practical definition of $\mathbf{F_h}$ as an explicit function of $\mathbf{f_h}$ and the observed parameter A_h for any cohorts studied and, for that matter, any form of cancer or other mortal disease.

Equations 21 and 23 alone do not yet define all three terms, $\mathbf{K_h}$, $\mathbf{F_h}$ and $\mathbf{f_h}$. There are three unknowns and only two independent equations. For the last necessary equation, one can take advantage of the feature that the mortality function OBS*(h,t) reaches a clear maximum in old age. The derivative of a continuous function equals zero at a maximum. By taking the derivative of OBS*(h,t) and setting it equal to zero at $t = t_{max,,}$ $\mathbf{F_h}$ can be written in terms of the other unknowns for any value of 'n'. Derivation of this general solution is shown as Equation 24, using a temporary function l(t) to represent the terms in the exponential of the probability of still being alive.

$$OBS^{*}(h,t) \sim \frac{F_h \cdot K_h \cdot (t - \Delta_h)^{n-1}}{F_h + (1 - F_h) \cdot e^{\frac{1}{f_h}\int_0^t (1 - S(h,t)) \cdot K_h \cdot (t - \Delta_h)^{n-1} dt}}$$

$$= \frac{F_h \cdot K_h \cdot (t - \Delta_h)^{n-1}}{F_h + (1 - F_h) \cdot e^{l(t)}}$$

Evaluating the derivative at age $t = t_{max}$, where the derivative of $OBS^{*}(h,t)$ equals 0:

$$\frac{d\,OBS^{*}(h,t)}{dt}\Big|_{t_{max}} = \frac{(n - 1) \cdot F_h \cdot K_h \cdot (t_{max} - \Delta_h)^{n-2} \cdot (F_h + (1 - F_h) \cdot e^{l(t_{max})})}{(F_h + (1 - F_h) \cdot e^{l(t_{max})})^2} -$$

$$\frac{F_h \cdot K_h \cdot (t_{max} - \Delta_h)^{n-1} \cdot \frac{d\,l(t)}{dt}\Big|_{t_{max}} \cdot (1 - F_h) \cdot e^{l(t_{max})}}{(F_h + (1 - F_h) \cdot e^{l(t_{max})})^2} = 0$$

Eliminating common terms this simplifies to:

$$(n - 1) \cdot (F_h \cdot e^{-l(t_{max})} + (1 - F_h)) - (t_{max} - \Delta_h) \cdot \frac{d\,l(t)}{dt}\Big|_{t_{max}} \cdot (1 - F_h) = 0$$

Evaluating the derivative of l(t):

$$(n - 1) \cdot f_h \cdot (F_h \cdot e^{-l(t_{max})} + (1 - F_h)) - (1 - S(h,t_{max})) \cdot K_h \cdot (t_{max} - \Delta_h)^{n} (1 - F_h) = 0$$

Solving for the fraction at risk, F_h, creates Equation 24:

$$F_h \cdot [(n - 1) \cdot f_h \cdot (1 - e^{-l(t_{max})}) - (1 - S(h,t_{max})) \cdot K_h \cdot (t_{max} - \Delta_h)^{n}] =$$

$$(n - 1) \cdot f_h - (1 - S(h,t_{max})) \cdot K_h \cdot (t_{max} - \Delta_h)^{n}$$

$$F_h = \frac{(n - 1) \cdot f_h - (1 - S(h,t_{max})) \cdot K_h \cdot (t_{max} - \Delta_h)^{n}}{(n - 1) \cdot f_h \cdot (1 - e^{-l(t_{max})}) - (1 - S(h,t_{max})) \cdot K_h \cdot (t_{max} - \Delta_h)^{n}}$$

(Eq. 24)

$$F_h = \frac{(n-1)\cdot f_h - (1 - S(h,t_{max}))\,\kappa_h\,(t_{max} - \Delta_h)^n}{(n-1)\cdot f_h \cdot \left(1 - e^{-\frac{1}{f_h}\int_0^{t_{max}} (1 - S(h,t))\,\kappa_h\,(t - \Delta_h)^{n-1}\,dt}\right) - (1 - S(h,t_{max}))\,\kappa_h\,(t_{max} - \Delta_h)^n}$$

This provides four independent equations using three separate features of the mortality curves, plus the direct observation of Δ_h:

1. the slope of the (n-1)th root of OBS*(h,t) for Equation 21,
2. the x-intercept of the (n-1)th root OBS*(h,t), direct estimation of Δ_h,
3. the area under $OBS^R(h,t)$ for Equation 23 and
4. the maximum of OBS*(h,t)) for Equation 24.

Together these equations allow the explicit determination of the two desired population risk parameters, $\mathbf{F_h}$ and $\mathbf{f_h}$ for any birth cohort for which these four features were defined by the data, which allows one to chart the health effects of environmental changes in populations. $\mathbf{F_h}$ is also the minimum value for G or E_h for any birth year cohort since when $E_h = 1$, $G = \mathbf{F_h}$ and vice versa. Both of these properties are of clear value in exploring the genetic and environmental interactions which lead to cancer.

The $\mathbf{f_h}$ approximation algebraically provides a solution for the limitation created by the ignorance of the causes of death that share primary risk factors with the OBServed cancer. The practical importance of $\mathbf{f_h}$ itself is however not entirely clear because it includes both historical

shifts in the accuracy of cancer diagnoses and in the probabilities of death by unknown connected diseases.

Furthermore, these equations explicitly derive a value for the physiological parameter K_h which can be used to estimate initiation mutation rates for any value of 'n' as shown in the following section.

3.5.5 Explicit Terms for Secondary Risk Parameters

3.5.5.1 The Product of Initiation Mutation Rates, $(r_i r_j r_k \ldots r_n)$.

Initiation with n>1 required events is modeled as per Armitage and Doll (1957) in which 'n' events in any cell would create the first cell of a precancerous lesion, extended to account the cell turnover in normal tissues and the organization of tissues as turnover units of constant and equal size containing N_a total cells at age 'a'.

The N_a total cells comprise terminal cells, transition cells and stem cells. The stem cell and each transition cell undergo τ divisions and no deaths per year. The terminal cells each "die" τ times per year and do not divide. The most probable pathway of initiation is therefore the accumulation of all but one of the 'n' mutations in the stem cell. The stem cell then repopulates its respective turnover unit with cells carrying the (n – 1) mutations, such that the nth mutation could now occur in any of these cells.

The rates of the required initiation mutations are represented as $\mathbf{r_i r_j r_k \ldots r_n}$, such that the expression describing the number of newly initiated cells in year 'a' is simply:

(Eq. 25)

Initiated cells during the year 'a' = $n \tau^n (\mathbf{r_i r_j r_k \ldots r_n}) N_a a^{n-1}$ (Armitage and Doll, 1954)

assuming that the order of the initiation mutations is not important.

3.5.5.2 Difference in Division and Death rates in Precancerous Lesions, $(\alpha-\beta)$, and Stochastic Extinction of Newly Initiated Cells.

As recognized and algebraically treated by Moolgavkar (1990b), each initiated cell could die before it divides. Even small colonies have a high probability that all cells will die; only a few would be expected to survive if the probability of cell division is only marginally greater than the probability of cell death. Given a cell division rate of α cell divisions per year and a death rate, β, for an initiated cell, the probability of non-extinction or survival is $(\alpha-\beta)/\alpha$ (Moolgavkar, 1990b).

The origin of this solution comes from the Gambler's Ruin problem. A gambler starting with 1 dollar (1 cell) makes 1 dollar bids and wins two dollars with probability $\alpha/(\alpha+\beta)$ (cell division) or loses the dollar with probability $\beta/(\alpha+\beta)$ (cell death). The probability that the gambler eventually loses all the money (stochastic extinction) is 1 minus the survival, or β/α.

Thus the number of newly arising and surviving precancerous lesions in year 'a' would be:

(Eq. 26)

$$\text{Surviving initiated precancerous lesions (a)} = \frac{(\alpha-\beta)}{\alpha}\, n\, \tau^n\, (\mathbf{r_i\, r_j\, r_k \ldots\, r_n})\, N_a\, a^{n-1}$$

All surviving precancerous lesions then have the property of inexorably giving rise to a lethal carcinoma via net growth and mutation.

The combination of the data of OBS*(h,t) and Equations 21, 23 and 24 allowed explicit determination of the unknown parameter $\mathbf{K_h}$ for any cohort studied. $\mathbf{K_h}$ is Nordling's annual rate of initiation per person modified to include Moolgavkar's necessary term for surviving stochastic extinction. For the case after N_a has reached a maximum, N_{max}, in young adults.

(Eq. 27)

$$\kappa_h = \frac{\alpha - \beta}{\alpha} n \tau^n (r_i r_j r_k \ldots r_n) N_{max}$$

Cell division and 'death' rates can be acquired from actual tissue samples: τ in normal tissue, and α and β in precancerous lesions. However, a value for the growth rate of the lesion, $(\alpha - \beta)$, is small and accurate independent estimates for the division and deaths rates *in vivo* are not plausible so as to properly estimate this difference from tissue samples.

One can make use of an interesting property of OBS*(h,t) to explicitly define $(\alpha-\beta)$ and then estimate the value of $(r_i r_j r_k \ldots r_n)$. For this, Nordling's model for the expected mortality from the OBServed disease, $P_{OBS}(h,t)$, needs to be extended to account for the growth rate of a precancerous lesion.

3.5.5.3 Probability of Promotion at Age 't' Given Initiation at Age 'a', m = 1

Assuming that the third stage of carcinogenesis, progression, occurs rapidly and can be effectively modeled as occurring in zero years, the expected mortality from the OBServed form of cancer simply equals the probability of initiation at age 'a' (Equation 26) times the probability of promotion occurring at a later age 't'. The model for the second stage of the three-stage carcinogenesis model, promotion, is again based on Armitage and Doll (1957) in which 'm' particular events in any cell of the precancerous lesion would create the first cell of a carcinoma. Since the number of promotion events needed for cancer is yet unknown, the simplest case to consider is that a single genetic event could turn an adenoma cell into a carcinoma cell.

Based on the Exponential distribution, the probability of at least one cell undergoing promotion at age 't' in an adenoma that was initiated at age 'a' is:

(Eq. 28)

$$\frac{d\left(1 - e^{\left(-r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{(2^{(\alpha-\beta)(t-a)} - 1)}{\ln 2} \cdot \frac{\alpha_c-\beta_c}{\alpha_c}\right)}\right)}{d(t-a)}$$

The product in the exponential represents the expected number of cancerous cells in the precancerous lesion that have been promoted (t – a) years after initiation and survival of stochastic extinction.

Here, r_A represents the promotion mutation rate per cell division, and $(\alpha_c - \beta_c) \div \alpha_c$ represents the probability of a promoted cell colony surviving stochastic extinction, given cell division and death rates per year of α_c and β_c respectively. The remaining terms in Equation 28 describe the total number of cell divisions, or chances for promotion, occurring within the precancerous lesion, derived as follows:

The initial number of cells in a lesion that has survived stochastic extinction, is not one, but $\alpha/(\alpha - \beta)$. (This is because of the stochastic redistribution of surviving cells among all initiated lesions into the surviving precancerous lesions, since stochastic processes cannot increase or decrease the total number of initiated cells in a population. This phenomenon has the effect of reducing the age-specific fraction of persons with an initiated colony but also considerably shortens the period expected to "promote" one cell in an initiated colony into a neoplastic cell for any number of required promotional events). The colony then grows at a doubling rate of $(\alpha - \beta)$ per year, such that the number of cells in the precancerous lesion after (t – a) years, the time since initiation, is:

$$\frac{\alpha}{\alpha-\beta} \cdot 2^{(t-a)(\alpha-\beta)}$$

Since the division and death rates of a precancerous lesion are approximately equal, a year can be divided into α periods, the chances per year for a precancerous cell to divide or die. The number of cells in the adenoma can be expressed as a function of the number of these periods that have elapsed, δ:

$$\frac{\alpha}{\alpha-\beta} \cdot 2^{[\alpha(t-a)]\frac{(\alpha-\beta)}{\alpha}} = \frac{\alpha}{\alpha-\beta} \cdot 2^{\delta\frac{(\alpha-\beta)}{\alpha}}$$

The number of total cell divisions having occurred in the adenoma can then be related to the number of cells. In order to have a colony of a certain size, N_{prec}, we recognize that there must have been $(N_{prec} \div 2)$ divisions **within** the last period of possible division:

$$\frac{\alpha}{\alpha-\beta} \cdot \frac{2^{\delta\frac{(\alpha-\beta)}{\alpha}}}{2}$$

Consequently, the total number of cell divisions is the sum of the number of divisions needed to give each of the intermediate sizes of the precancerous lesion up to the last period, δ:

$$\sum_{i=1}^{\delta} \frac{\alpha}{\alpha-\beta} \cdot \frac{2^{i\frac{(\alpha-\beta)}{\alpha}}}{2} = \sum_{i=1}^{\alpha(t-a)} \frac{\alpha}{\alpha-\beta} \cdot \frac{2^{i\frac{(\alpha-\beta)}{\alpha}}}{2}$$

This summation can be solved explicitly as:

$$\frac{\alpha}{\alpha-\beta} \cdot \frac{2^{-\beta/\alpha}\left(2^{(\alpha-\beta)(t-a)} - 1\right)}{2^{(\alpha-\beta)/\alpha} - 1}$$

Taking the integral instead of the summation above gives a reliable and easier to remember estimate:

$$\left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{(2^{(\alpha-\beta)(t-a)} - 1)}{2 \ln 2}$$

Last, for every division, each of the two daughter cells can acquire the promotion mutation. Therefore, the number of opportunities for promotion after (t – a) years is just twice the total number of divisions:

$$\left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{(2^{(\alpha-\beta)(t-a)} - 1)}{\ln 2}$$

as included in Equation 28.

3.5.5.4 Probability of death at age 't' given individual is at risk for cancer, $\mathbf{P_{OBS}(h,t)}$

Combining the probability of initiation from Equation 25 with the probability of promotion from Equation 28, one gets that the expected probability of death at age 't' for an individual at risk is (m = 1 case illustrated):

(Eq. 29)

$$P_{OBS}(h,t) = n\tau^{n} r_i r_j r_k \ldots r_n \frac{\alpha-\beta}{\alpha} \int_0^t a^{n-1} N_a \frac{d\left(1 - e^{\left(-r_A \left(\frac{\alpha}{\alpha-\beta}\right)^2 \frac{(2^{(\alpha-\beta)(t-a)} - 1)}{\ln 2} \frac{\alpha_c - \beta_c}{\alpha_c}\right)}\right)}{d(t-a)} da$$

Obviously, an individual could develop more than one precancerous lesion within a lifetime. In its simplest form, $\mathbf{P_{OBS}(h.t)}$ is then just the probability of a cell being initiated at any age 'a' and having any one of its descendants promoted to a carcinoma at age 't', expressed as the convolution of the probabilities of initiation and promotion.

3.5.5.5 Explicit determination of the growth rate of precancerous lesions

The growth rate of precancerous lesions can be estimated directly from the mortality curves, derived as Equation 30 below. (Although illustrated for case n=2, m=1, the calculated adenomatous growth rate is approximately the same for all other cases.)

For ages below Δ_h, OBS*(h,t) is approximately OBS*(h,t) = OBS(h,t) ÷ [R(h,t) (1 – S(h,t))] ≈ $\mathbf{F_h}$ $P_{OBS}(h,t)$, since most individuals at risk are not expected to have already died of either the observed form of death or anyone of the connected forms of death. As a first approximation, assume a constant number of cells in the target tissue. For clarity, parameters of Equation 29 that do not vary with age have been grouped (e.g. $\mathbf{F_h}$ is included in C1):

$$OBS^{*}(h,t) \approx C_1 \int_0^t a \frac{d\left(1 - e^{-C_2 \cdot (2^{(\alpha-\beta)(t-a)} - 1)}\right)}{d(t-a)} da$$

To help solve this integral, one can approximate $e^x \approx 1 + x$ when x is small. This yields:

$$OBS^{*}(h,t) \approx C_1 \int_0^t a \frac{d\left(1 - (1 - C_2 \cdot (2^{(\alpha-\beta)(t-a)} - 1))\right)}{d(t-a)} da \approx C_1 C_2 \int_0^t a \frac{d(2^{(\alpha-\beta)(t-a)} - 1))}{d(t-a)} da$$

$$\approx C \int_0^t a [2^{(\alpha-\beta)(t-a)} (\alpha-\beta)\ \ln 2]\, da \approx \frac{C}{(\alpha-\beta)\ \ln 2} [2^{(\alpha-\beta)t} - t\,(\alpha-\beta)\ln 2 - 1]$$

combining the two constants, $C_1C_2 = C$. Taking the derivative of OBS*(h,t):

$$\frac{d(OBS^{*}(h,t))}{dt} \approx C\,[2^{(\alpha-\beta)t} - 1] \approx C\,2^{(\alpha-\beta)t}$$

The $\log_2$ of dOBS*(h,t) ÷ dt is therefore a function of 't' whose slope is the growth rate of the precancerous lesion, $\alpha - \beta$:

(Eq. 30)

$$\log_2 \frac{d(OBS^{*}(h,t))}{dt} \approx (\alpha-\beta)t + \log_2(C)$$

This approximation is valid only when $2^{(\alpha-\beta)t} >> 1$, so when estimating the adenomatous growth rate, one needs to be careful not to use data from the age groups below age 18.

Fig. 41: Determination of $(\alpha - \beta)$ from the slope of $\log_2 \Delta(OBS^*(h,t)) \div \Delta t$.

(Data is for colon cancer of EAM born in the 1920s, ages 17.5 to 57.5)


Age (years)
10
20
30
40
50
60
-10
-15
-20
-25
log2Δ(OBS*(h,t)) ÷ Δt

If the increase in cell number during childhood were accounted for, the derivative of OBS*(h,t) for t > 17.5:

$$\frac{d(OBS^{*}(h,t))}{dt} \approx C\left[2^{(\alpha-\beta)t} - \frac{2^{16.5(\alpha-\beta)}}{2-2^{-16.5(\alpha-\beta)}(1+16.5\ln 2(\alpha-\beta)}\right] \approx C\,2^{(\alpha-\beta)t}$$

The $\log_2$ of this function plotted vs. t is indeed a straight line for ages for which the probability of promotion is yet small (i.e. $t < \Delta_h$), whose slope provides an estimate for the cell kinetic growth rate of the precancerous lesions, $(\alpha - \beta)$. As an example, Figure 41 shows an estimate for EAM born in the 1920s. To evaluate the derivative of OBS*(h,t) from the mortality data, the approximation $\Delta(OBS^{*}(h,t)) \div \Delta t \approx d(OBS^{*}(h,t)) \div dt$ is used.

3.5.5.6 Explicit Determination of the Product of Initiation Mutation Rates

The determination of the cell growth rate of a precancerous lesion, $(\alpha - \beta)$, allows us to calculate the product of the initiation mutation rates, $(r_i r_j r_k \ldots r_n)$ using the previously derived value of K_h in Equation 27. From this product, the geometric mean can be derived as $(r_i r_j r_k \ldots r_n)^{(1/n)}$.

3.5.5.7 Explicit Determination of the Promotion Mutation Rate for the Case of m = 1

Evaluation of the average promotion mutation rate can be expressed in terms of previously defined values. The value Δ_h represented the average time between initiation and promotion. The cumulative probability of promotion in a precancerous lesion is therefore approximately one-half, Δ_h years after its initiation. Using Equation 28, this means:

(Eq. 31)

$$\frac{1}{2} = 1 - e^{-r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{1}{\ln 2} \cdot \frac{\alpha_c-\beta_c}{\alpha_c} \cdot (2^{(\alpha-\beta)\Delta_h} - 1)}$$

$$\Delta_h = \frac{\log_2\left[1 + \left(r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{1}{[\ln(2)]^2} \cdot \frac{\alpha_c-\beta_c}{\alpha_c}\right)^{-1}\right]}{\alpha-\beta}$$

However, the assumption that the cumulative probability is approximately one-half at the average time between initiation and promotion is good only if the distribution for the probability of promotion can be approximated by a normal distribution. As $(\alpha - \beta)$ increases, deviation from normality occurs. If this is the case, the exact solution is essential.

For a continuous random variable, i.e. the time between initiation and promotion, t – a, can be defined as:

$$\Delta_h = \int_{-\infty}^{\infty} (t - a)\, P[t - a]\, d(t-a)$$

The expected time between initiation and promotion is then:

(Eq. 32)

$$\Delta_h = \int_0^{\infty} (t-a) \frac{d(1 - e^{-r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{1}{\ln 2} \cdot \frac{\alpha_c-\beta_c}{\alpha_c} \cdot (2^{(\alpha-\beta)(t-a)} - 1)})}{d(t-a)} d(t-a)$$

$$= \frac{-e^{-r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{1}{\ln 2} \cdot \frac{\alpha_c-\beta_c}{\alpha_c}} \cdot \mathrm{Ei}[r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{1}{\ln 2} \cdot \frac{\alpha_c-\beta_c}]}{(\alpha-\beta)\ln 2}$$

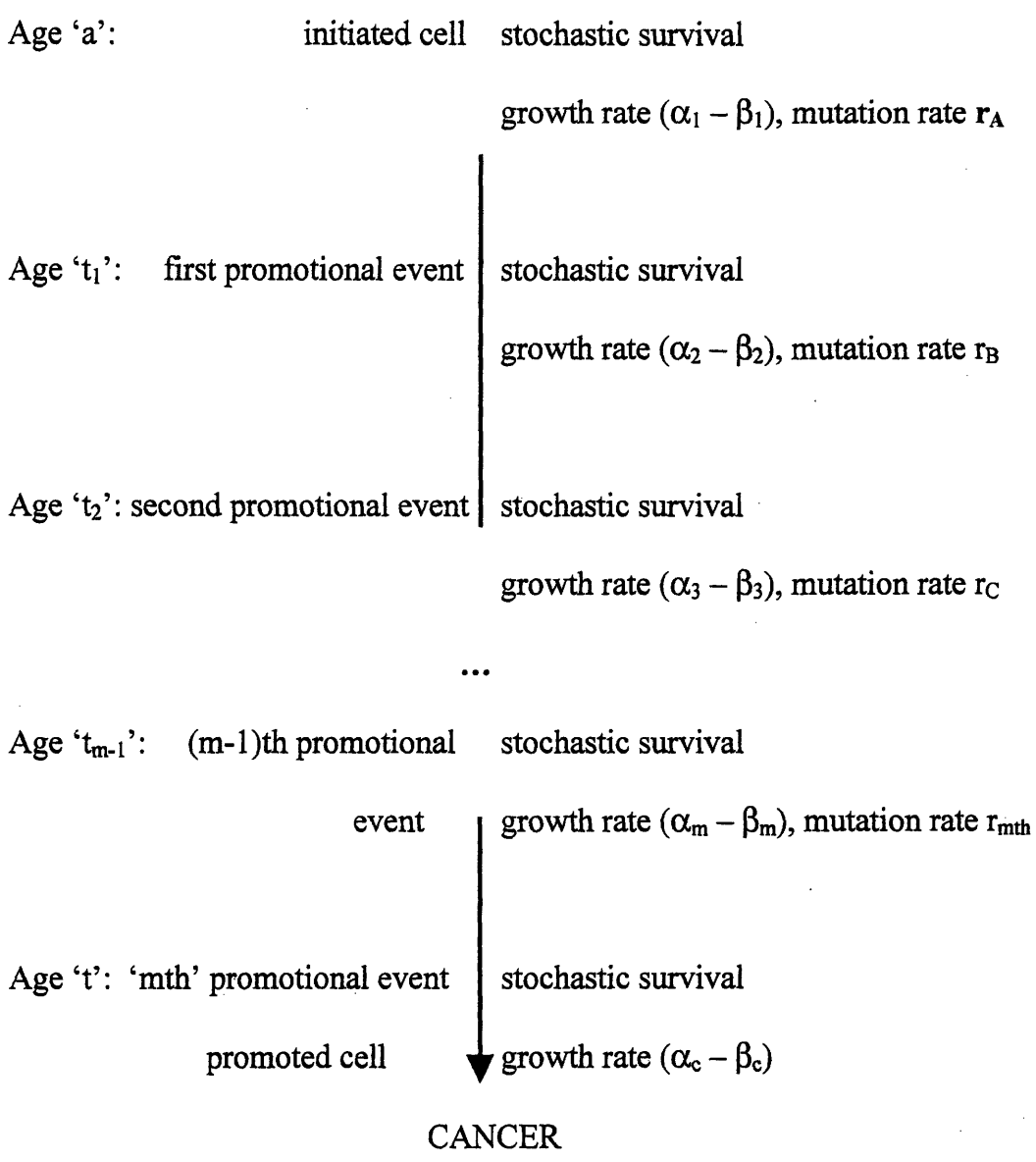
where Ei is the exponential integral function. A computational tool such as Mathematica™ (ExpIntegralEi function) or Matlab™ (expint function) can be used to evaluate the exponential integral function.

The only unknown parameter is the promotion mutation rate, $\mathbf{r_A}$. Either of these equations is sufficient to evaluate this last unknown parameter, thereby completing our explicit derivation of all physiological parameters in the three-stage carcinogenesis model, $\mathbf{r_i}$, $(\alpha - \beta)$ and $\mathbf{r_A}$.

3.5.5.8 Explicit Determination of the Promotion Mutation Rate for the Case of $m > 1$

If more than one genetic event were needed to convert a precancerous cell into a cancerous one, $m > 1$, when a cell in the precancerous lesion acquires the first promotion mutation, this cell has the potential to divide and become a distinct colony of cells now containing one promotional mutation. A cell within this colony is a target for a second promotional event, producing a new colony within the precancerous lesion, now made up of cells with two promotional mutations. This process continues until a cell acquires all necessary 'm' promotional mutations, thereby producing a carcinoma cell. As was the case for a newly initiated cell, any cell that has acquired a new promotion mutation could undergo stochastic extinction before developing a colony.

The precancerous lesion itself would thus appear to be a mix of colonies of cells containing zero or more of the promotional events, and the delay in the rise of the mortality curves, Δ_h, is now the sum of the average time between each promotional event.



The probability of promotion <u>at</u> age 't' simply follows as:

Probability of 1st promotion mutation by age ($a < t_1 < t$) x

Probability of 2nd promotion mutation by age ($t_1 < t_2 < t$) x

... x

Probability of (m-1)th promotion mutation by age ($t_{m-2} < t_{m-1} < t$) x

Probability of m^{th} promotion mutation **<u>at</u>** age t

Explicitly, this is:

$$\int_{a}^{t} \frac{d\left(1 - e^{\left(-mr_A \cdot \left(\frac{\alpha_1}{\alpha_1-\beta_1}\right)^2 \cdot \frac{2^{(\alpha_1-\beta_1)(t_1-a)} - 1}{\ln 2} \cdot \frac{\alpha_2-\beta_2}{\alpha_2}\right)}\right)}{d(t_1 - a)}$$

$$\int_{t_1}^{t} \frac{d\left(1 - e^{\left(-(m-1)r_B \cdot \left(\frac{\alpha_2}{\alpha_2-\beta_2}\right)^2 \cdot \frac{2^{(\alpha_2-\beta_2)(t_2-t_1)} - 1}{\ln 2} \cdot \frac{\alpha_3-\beta_3}{\alpha_3}\right)}\right)}{d(t_2 - t_1)}$$

$$\cdots$$

$$\int_{t_{m-2}}^{t} \frac{d\left(1 - e^{\left(-2r_{m-1} \cdot \left(\frac{\alpha_{m-1}}{\alpha_{m-1}-\beta_{m-1}}\right)^2 \cdot \frac{2^{(\alpha_{m-1}-\beta_{m-1})(t_{m-1}-t_{m-2})} - 1}{\ln 2} \cdot \frac{\alpha_m-\beta_m}{\alpha_m}\right)}\right)}{d(t_{m-1} - t_{m-2})}$$

$$\frac{d\left(1 - e^{\left(-r_m \cdot \left(\frac{\alpha_m}{\alpha_m-\beta_m}\right)^2 \cdot \frac{2^{(\alpha_m-\beta_m)(t_{m-1}-t_{m-2})} - 1}{\ln 2} \cdot \frac{\alpha_c-\beta_c}{\alpha_c}\right)}\right)}{d(t - t_{m-1})}$$

$$dt_1\, dt_2\, dt_3 \ldots dt_{m-1}$$

where the first promotion mutation occurs at any age, t_1, between initiation and death, the second promotion mutation occurs at any age, t_2, between the first mutation and death, the third promotion mutation occurs at any age, t_3, between the second mutation and death, and so forth until the last promotion mutation which must occur at age t, death. As was the case for initiation, there are 'm' target alleles for the first promotion event, (m - 1) for the second promotion event, and so forth.

Supposing that the acquisition of a new promotional event alters either the cell kinetic rates, α_x and β_x, or the promotion mutation rate, r_x, for that cell, it is not possible to estimate each cell kinetic rate and promotion mutation rate independently for each step. However, there exists an average promotion mutation rate, $\mathbf{r_A}$, and an average cell kinetic growth rate, $\alpha - \beta$that describes a similar process of 'm' promotional mutations such that the total delay of onset of the disease is the same. The probability of promotion is now of the 'simpler' form:

$$\int_{a}^{t} \frac{d\left(1 - e^{\left(-m\mathbf{r_A} \cdot \frac{\alpha}{\alpha-\beta} \cdot \frac{2^{(\alpha-\beta)(t_1-a)} - 1}{\ln 2}\right)}\right)}{d(t_1 - a)}$$

$$\int_{t_1}^{t} \frac{d\left(1 - e^{\left(-(m-1)\mathbf{r_A} \cdot \frac{\alpha}{\alpha-\beta} \cdot \frac{2^{(\alpha-\beta)(t_2-t_1)} - 1}{\ln 2}\right)}\right)}{d(t_2 - t_1)}$$

$$\cdots$$

$$\int_{t_{m-2}}^{t} \frac{d\left(1 - e^{\left(-2\mathbf{r_A} \cdot \frac{\alpha}{\alpha-\beta} \cdot \frac{2^{(\alpha-\beta)(t_{m-1}-t_{m-2})} - 1}{\ln 2}\right)}\right)}{d(t_{m-1} - t_{m-2})}$$

$$\frac{d\left(1 - e^{\left(-\mathbf{r_A} \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2 \cdot \frac{2^{(\alpha-\beta)(t-t_{m-1})} - 1}{\ln 2} \cdot \frac{\alpha_c-\beta_c}{\alpha_c}\right)}\right)}{d(t - t_{m-1})} dt_1\, dt_2\, dt_3 \ldots dt_{m-1}$$

The estimate of the average promotion mutation rate can be written with respect to the average interarrival time Δ_x between each promotional event. Using the approximation that the

average time approximately corresponds to the time when the cumulative probability for that promotion mutation is 0.5:

$$\Delta_1 = \frac{\log_2\left(1+\left[m r_A \cdot \frac{\alpha}{\alpha-\beta}\cdot\frac{1}{[\ln(2)]^2}\right]^{-1}\right)}{\alpha-\beta}, \quad \Delta_2 = \frac{\log_2\left(1+\left[(m-1) r_A \cdot \frac{\alpha}{\alpha-\beta}\cdot\frac{1}{[\ln(2)]^2}\right]^{-1}\right)}{\alpha-\beta}$$

• • •

$$\Delta_{m-1} = \frac{\log_2\left(1+\left[2 r_A \cdot \frac{\alpha}{\alpha-\beta}\cdot\frac{1}{[\ln(2)]^2}\right]^{-1}\right)}{\alpha-\beta}, \quad \Delta_m = \frac{\log_2\left(1+\left[r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2\cdot\frac{1}{[\ln(2)]^2}\cdot\frac{\alpha_c-\beta_c}{\alpha_c}\right]^{-1}\right)}{\alpha-\beta}$$

The total expected delay between initiation and promotion, Δ_h, is simply the sum of the delays between each promotion mutation. For m>1:

(Eq. 33)

$$\Delta_h = \frac{\sum_{i=2}^{m}\log_2\left(1+\left[i r_A \cdot \frac{\alpha}{\alpha-\beta}\cdot\frac{1}{[\ln(2)]^2}\right]^{-1}\right)}{\alpha-\beta} + \frac{\log_2\left(1+\left[r_A \cdot \left(\frac{\alpha}{\alpha-\beta}\right)^2\cdot\frac{1}{[\ln(2)]^2}\cdot\frac{\alpha_c-\beta_c}{\alpha_c}\right]^{-1}\right)}{\alpha-\beta}$$

assuming that the order the promotion mutations occur is inconsequential.

Supposing instead that each of the promotion mutations leads to either an elevated cell growth rate or an elevated mutation rate per cell year, then there might exist a particular order for the 'm' necessary promotion mutations that is most favorable for promotion of the tumor. If these deleterious promotion mutations do not occur early in the order of the 'm' mutations, the individual might not accumulate all of the necessary promotion mutations within their lifetime.

Using the same logic as above, we can explicitly evaluate the delay between initiation and promotion, if the 'm' mutations had to occur in a particular order:

(Eq. 34)

$$\Delta_{\mathbf{h}} = \frac{(m-1)\cdot \log_2\left(1+\left[r_A \cdot \frac{\alpha}{\alpha-\beta}\cdot\frac{1}{[\ln(2)]^2}\right]^{-1}\right)}{\alpha-\beta} + \frac{\log_2\left(1+\left[r_A\cdot\left(\frac{\alpha}{\alpha-\beta}\right)^2\cdot\frac{1}{[\ln(2)]^2}\cdot\frac{\alpha_c-\beta_c}{\alpha_c}\right]^{-1}\right)}{\alpha-\beta}$$

Equations 33 and 34 allow one to estimate the average promotion mutation rate for the case where the 'm' promotion mutations occur in a completely unordered manner and the case where the 'm' promotion mutations must follow in a particular order. Of course, it is possible that only some of the promotion mutations must occur in order. For this case, where the 'm' mutations are only partially ordered, these equations would describe the possible range for the average promotion mutation rate.

In the process above, the average promotion mutation rate and the average growth rate of the precancerous lesion refer to the average rate of only the precancerous cells that comprise the direct lineage between the first precancerous cell and the first cancerous one.

3.6 COMPUTER APPLICATIONS

3.6.1 Calculating the Expected Mortality Rate Given a Set of Values for $\mathbf{F_h}$, $\mathbf{f_h}$, $\mathbf{r_i}$, $(\alpha - \beta)$, and $\mathbf{r_A}$ - Mathematica™

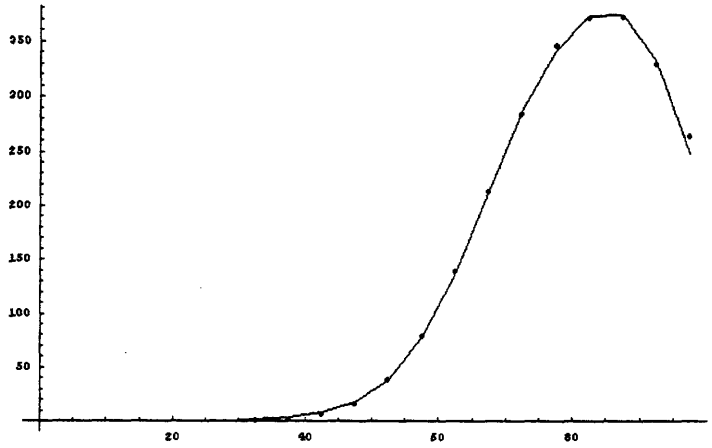
The main advantage of the Mathematica™ software package is that allows a user to write code for a mathematical model as is written in the derived Equations above. Included herein is

the code for the case of 2 initiation mutations and 1 promotion mutation being necessary for carcinogenesis. For demonstration, the parameters that have been plugged in to this program are the maximum fit solution for the 1890 EAM lung cancer mortality data, assuming that there is only one subpopulation at risk, the smokers. These are not the parameters listed in the Results section which do correct for the fraction of individuals who died of lung cancer by means other than smoking. Comments are enclosed in the code as (* COMMENT *):

```
(* Observed mortality data: Lung cancer 1890s EAM {age,rate} *)
DataOBS = {{32.5, 0.89}, {37.5, 2.32}, {42.5, 6.10}, {47.5, 16.22}, {52.5, 38.08},
  {57.5, 78.86}, {62.5, 139.38}, {67.5, 212.90}, {72.5, 283.78}, {77.5, 345.98},
  {82.5, 370.73}, {87.5, 371.51}, {92.5, 329.36}, {97.5, 264.36}};

(* Gender: 0 for females, 1 for males *)

boys = 1;

(* Gender dependent growth rate of child between ages 0 and 1.5 years old *)
(* 1st term (==) checks for equality;
 2nd term sets ξ = 1.23 if true; 3rd term sets ξ = 1.17 if false *)

ξ = If[boys == 0, 1.23, 1.17];

(* Gender dependent growth rate after age 1.5 *)

η = If[boys == 0, 0.167, 0.159];

(* Age at end of growth *)

puberty = If[boys == 0, 14.5, 16.5];

(* Number of cells at end of growth *)

N_max = 2.4 10^8;

(* Number of cells at age 1.5 years old *)

N_1.5 = N_max / 2^η(puberty-1.5);
```

```
(* Turnover rate of normal tissue *)

τ = 5.7;

(* Rate of 1st initiation mutation *)

r_i = 0.039 10^-5;

(* Rate of 2nd initiation mutation;
 3 times as large as first *)

r_j = 3 r_i;

(* Division rate of precancerous cells *)

α = 12.7;

(* Death rate of precancerous cells *)

β = α - 0.243;

(* Rate of promotion mutation *)

r_A = 0.0155 10^-5;

(* Division rate of cancerous cells *)

α_C = 42.2;

(* Death rate of cancerous cells *)

β_C = 28;

(* Fraction at risk *)

F = 0.9617;

(* Relative risk *)

f = 0.0418;

(* Survival rates *)
(* This is the construct for a function named S, whose variable is t *)


S[t_] := 0;

(* So that the program does not need to continuously calculate the derivative
  term for the probability of promotion, (1-E^(-r_A (α/(α-β))^2 (2^((α-β)(t-a))-1) (α_C-β_C)/α_C)),
 it is advisable to have this precalculated as its own function *)

Promotion[t_] := 2^((-a+t)(α-β)) e^((-1+2^((-a+t)(α-β))) α^2 r_A (α_C-β_C))/((α-β)^2 α_C) α^2 Log[2] r_A (α_C - β_C)/((α - β) α_C);
```

```
(* Function PS(t) = P_OBS(h,t) times (1 - S(h,t)) *)
(* The extra PS[t]= term tells the computer to memorize the calculated
   value for a given t once it has already being calculated;
   this helps speed up integration of PS(t) in the next line; *)
 (* The Clear command is so that whenever new parameters are used,
 the computer will clear memory *)
 (* The If statement is to account for the number of cells
  at the age of initiation; the integrals (NIntegrate) are cut
  into the three age periods of when initiation could occur*)

Clear[PS]
PS[t_] := PS[t] =
   2 τ^2 r_i r_j N_max (α - β)/α
    If[t ≤ 1.5, NIntegrate[a N_1.5/N_max 2^γ(a-1.5) Promotion[t], {a, 0, 1.5}],
     If[t ≤ puberty, NIntegrate[a N_1.5/N_max 2^γ(a-1.5) Promotion[t], {a, 0, 1.5}] +
                NIntegrate[a 2^η(a-puberty) Promotion[t], {a, 1.5, puberty}],
      NIntegrate[a N_1.5/N_max 2^γ(a-1.5) Promotion[t], {a, 0, 1.5}] +
                NIntegrate[a 2^η(a-puberty) Promotion[t], {a, 1.5, puberty}] +
                NIntegrate[a Promotion[t], {a, puberty, t}]]] (1 - S[t]);

(* OBServed mortality rate per 100,000*)

OBS[t_] := OBS[t] = F PS[t] / ((1 - S[t]) (F + (1 - F) E^(1/f NIntegrate[PS[u], {u, 0, t}]))) 10^5;

(* Calculates expected mortality
  rate given set of parameters for ages 2.5,7.5,...,97.5 *)
DataModel = Table[{t, OBS[t]}, {t, 2.5, 97.5, 5}];

(* Plots observed data as points; plots model results as line *)
Show[ListPlot[DataOBS], ListPlot[DataModel, PlotJoined -> True]]
```



3.6.2 Calculating the expected mortality rate given a set of values for $\mathbf{F_h}$, $\mathbf{f_h}$, $\mathbf{r_i}$, $(\alpha - \beta)$, and $\mathbf{r_A}$ - Matlab™

For each set of parameters, the Mathematica™ program takes approximately 30-45 seconds (Pentium III, 500 MHz) to calculate the expected mortality rate; the lengthy calculation is primarily due to the fact that the integral term cannot be explicitly solved; as such, numerical integration (NIntegrate) had been used. To maximize the fit of the solution requires multiple iterations, such that implementation of the model would require a different speedier software package. The above program is nonetheless indispensable so as to verify the results of any other programs.

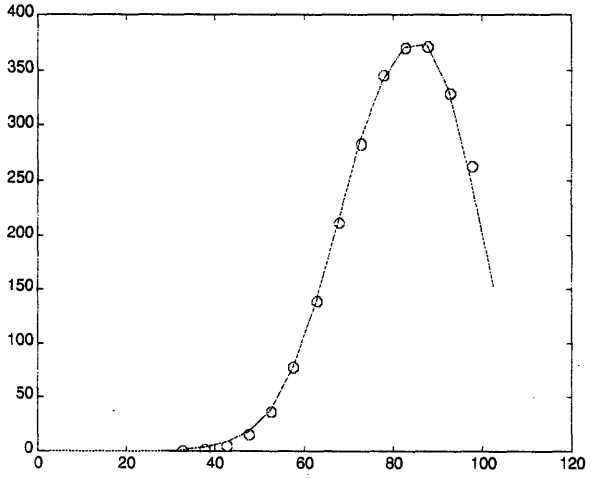
The Mathematica™ code was translated into Matlab™. Comments now are shown as % COMMENT. The code is included herein for documenting purposes.

```
%Observed mortality data: Lung cancer 1890s EAM
age = 32.5:5:97.5;
dataOBS = [0.89
2.32
6.10
16.22
38.08
78.86
139.38
212.90
283.78
345.98
370.73
371.51
329.36
264.36]';
boys = 1;

% Gender: 0 for females, 1 for males
if boys == 0            % Rates for females (==) checks for equality
   stopgrowth = 14.5;   % Age at end of growth
   zeta = 1.23;         % Growth rate of child between ages 0 and 1.5
   eta = 0.167;         % Growth rate after age 1.5
else
   stopgrowth = 16.5;   % Rates for males
   zeta = 1.17;
   eta = 0.159;
```

```
end

% Number of cells in adulthood
Nmax = 2.4*10^8;

% Number of cells at age 1.5 years old
Nchild = Nmax/2^(zeta*(stopgrowth-1.5));

% Turnover rate of normal tissue
tau = 5.7;

% Rate of 1st initiation mutation
ri = 0.039*10^(-5);

% Rate of 2nd initiation mutation; 3 times as large as first
rj = 3*ri;

% Division rate of precancerous cells
alpha = 12.7;

% Death rate of precancerous cells
beta = alpha - 0.243;

% Rate of promotion mutation
rA = 0.0155*10^(-5);

% Division rate of cancerous cells
alphac = 42.2;

% Death rate of cancerous cells
betac = 28;

% Fraction at risk
F = 0.9617;

% Relative risk
f = 0.0418;

% Survival rates for all ages 0.5,1.5,...,102.5
% in initializing an array, the '( ,1)' informs it only one column of values
S = zeros(103,1);

% Hazard function PS(t) = POBS(h,t) / (1 - S(h,t))
% Hazard function for cells initiated during the first 1.5 years of age
% have been excluded as they were verified to have negligible contribution
% when code in Mathematica program for these adenomas was removed
PS = zeros(103,1);

for t=2.5:102.5                        % t - age at death
   h=min(t,stopgrowth);                % Contribution of lesions prior to
   integx=[1.5:h]                      % reaching adulthood; for ages at death
   integy=zeros(length(integx),1); % less than adulthood, we consider only
                                       % lesions initiated prior to t (min)
   for a=1.5:h                         % a - age at initiation
      integy((a-1.5)+1)=2*tau^2*ri*rj*Nmax*(alpha-beta)/alpha*a*2^(eta*(a-
stopgrowth))*2^((alpha-beta)*(t-a))*exp(-rA*(alpha/(alpha-
```

```
beta))^2*(2^((alpha-beta)*(t-a))-1)*(alphac-betac)/alphac)*alpha^2/(alpha-
beta)*rA*(alphac-betac)/alphac*log(2)*(1-S(t+0.5));
   end
   % integy - calculates contribution of lesion initiated at age a to death
   % at age t; for loop is used to calculate for all ages a. Note that it is
   % the same calculation as Mathematica program prior to integration

   PS(t+0.5) = PS(t+0.5)+trapz(integy);
   % integration by trapezoids; similar to NIntegrate in Mathematica program
   % Note, arrays can only take integers, so we add 0.5 to ages to
   % correspond to array position
end

% same as previous loop, but now we consider only lesions initiated at ages
% between adulthood and death
for t=stopgrowth+1:102.5
   integx=[stopgrowth:t];
   integy=zeros(length(integx),1);
   for a=stopgrowth:t
      integy((a-stopgrowth)+1)=2*tau^2*ri*rj*Nmax*(alpha-
beta)/alpha*a*2^((alpha-beta)*(t-a))*exp(-rA*(alpha/(alpha-
beta))^2*(2^((alpha-beta)*(t-a))-1)*(alphac-betac)/alphac)*alpha^2/(alpha-
beta)*rA*(alphac-betac)/alphac*log(2)*(1-S(t+0.5));
   end
   PS(t+0.5) = PS(t+0.5)+trapz(integy);
end

% We now calculate the mortality rate, OBS(h,t) for ages 2.5, 7.5,..., 102.5
% There are 21 values; to convert from age to array position, we simply add
% 2.5 to the age and divide by 5 (i.e. (2.5+2.5)/5 = 1)
OBS=zeros(21,1);
for t=2.5:5:102.5
   OBS((t+2.5/5)) = F*PS(t+0.5)/((1-S(t+0.5))*(F+(1-
F)*exp(trapz(PS(2.5+0.5:t+0.5))/f)))*10^5;
End

% plot observed as points; plot model as line
plot(age,dataOBS,'o',2.5:5:102.5,OBS)
```

Output:



3.6.3 Calculating the expected mortality rate given a set of values for $\mathbf{F_h}$, $\mathbf{f_h}$, $\mathbf{r_i}$, $(\alpha - \beta)$, and $\mathbf{r_A}$ - Microsoft Excel™ Template

Although faster than its Mathematica™ predecessor, the Matlab™ program could not be effectively used for maximum likelihood routines. However, it does demonstrate that using integration by trapezoids gives results that are comparable to the exact calculations made by the Mathematica™ program.

Due to the existence of a built-in routine for maximum likelihood in Microsoft Excel™ and the limited loss in accuracy in integration by trapezoids, we finally opted to develop a template in Microsoft Excel™ which not only can calculate expected mortality rates for a given set of parameters, but can also adjust these parameters until the expected mortality rates match those observed. The template consists of two worksheets, one containing the Raw Data, the other containing the formulas for calculating mortality rates and the Fitting routine.

Figure 42 illustrates the first worksheet (Raw Data) which contains the observed data for the age-specific mortality rates for a birth decade cohort of interest corrected for underreporting, $OBS^R(h,t)$. The column labeled Weight refers to how to weigh the fitting of each point. The following are the weighing factors that can be used depending on which points are to be best fit:

Population values/100,000	-	Chi-square (favors early age points)
1	-	Sum of the differences squared of observed versus expected mortality rates, divided by the expected mortality rate (favors points with lower mortality rates)
0	-	Tells maximum likelihood routine to ignore point

Expected number of deaths - Sum of the differences squared of the number of observed versus expected deaths (favors points with the largest number of deaths)

Only this last method requires a formula (as shown in Figure 42). The formulae refer to cells in column HG of the second worksheet (Fitting) corresponding to the expected mortality rates, as shown in Figure 48. (We divide by 100,000 to get the true mortality since the observed mortality data had been expressed as per 100,000).

Figures 43-51 illustrate the second worksheet (Fitting) that calculates expected mortality rates for a given set of parameters. For instructional purposes, the template is shown in duplicate form, with and without revealing formulas so that the user can distinguish which terms must be input and which terms are calculated by the template. (To have Excel™ exhibit a formula instead of the values to the formula, go to the Tools Menu, and click on Options; under the View panel, there is a click box to have the worksheets show formulae or not).

Instructions on how to use this template are included in the legend to the figures. [3]

[3] The construction of the template is by no means easy. Learning about the 'filling' feature in Excel™ to speed the construction is recommended, as this feature helps fill multiple formulae with similar patterns. Integration has been approximated by trapezoids every 0.5 years. To improve upon the accuracy of the template, instead of calculating the contribution of precancerous lesions initiated every 0.5 years, one could calculate contributions in smaller time intervals. Given that the results, as shown in Figure 43, are similar to those of the Mathematica™ program, we have not extended our template to do so.

Fig. 42: MAXIMUM LIKELIHOOD TEMPLATE: Raw Data worksheet

Instructions:

1) Transcribe mortality data adjusted for underreporting, $OBS^{R}(h,t)$, into Column B, by appropriate age category.

2) If using Chi-Square or the least sum of errors of the number of deaths to fit mortality data, include population values in Column D.

3) Type in which weighing factor to use for maximum likelihood. (Weighing factors are described in Section 3.6.3) Illustrated is the case using the least sum of errors of the number of deaths as the maximum likelihood method to fit the data; reference 'Fitting!HG' is to the calculated expected mortality rate as shown in Figure 49). The number references 11, 20, 30,...,210 are the row number corresponding to the appropriate age groups (See Figure 49).

4) If using Chi-Square to fit mortality data, the weighing factor is the population divided by 100,000:

   i.e. for the age group of 3, cell C2 should read: =D2/10^5

   D2 refers to the population parameter in cell D2, and since mortality data had been expressed as per 100,000 we adjust this back to the true rate.

Microsoft Excel - N2M1

File Edit View Insert Format Tools Data Window Help

	A	B	C	D
1	Age	Data	Weight	Population
2	3		0	
3	7.5		0	
4	12.5		0	
5	17.5		0	
6	22.5		0	
7	27.5		0	
8	32.5	0.892303588	116.4209088	8,060,493
9	37.5	2.318752323	1001.231748	29,309,253
10	42.5	6.102864124	3138.142283	39,673,183
11	47.5	16.22427966	7072.44374	39,608,381
12	52.5	38.07965624	14652.5844	37,967,474
13	57.5	78.86345626	28042.52126	36,305,238
14	62.5	139.3796775	45423.07149	32,957,586
15	67.5	212.8998573	60714.47184	28,355,398
16	72.5	283.7833627	64957.58147	22,605,333
17	77.5	345.9750636	55468.98629	16,280,034
18	82.5	370.7324432	38339.98897	10,309,444
19	87.5	371.5141656	20023.51967	5,368,021
20	92.5	329.3603306	6954.271556	2,099,980
21	97.5	264.3553957	628.2306154	253,614
22	102.5		0	

Raw Data | Fitting

Microsoft Excel - N2M1 Formulas

File Edit View Insert Format Tools Data Window Help

	A	B	C	D
1	Age	Data	Weight	Population
2	3		=D2/10^5*Fitting!HG11	
3	7.5		=D3/10^5*Fitting!HG20	
4	12.5		=D4/10^5*Fitting!HG30	
5	17.5		=D5/10^5*Fitting!HG40	
6	22.5		=D6/10^5*Fitting!HG50	
7	27.5		=D7/10^5*Fitting!HG60	
8	32.5	0.892303587616973	=D8/10^5*Fitting!HG70	8060493
9	37.5	2.31875232348056	=D9/10^5*Fitting!HG80	29309253
10	42.5	6.10286412408995	=D10/10^5*Fitting!HG90	39673183
11	47.5	16.2242796553296	=D11/10^5*Fitting!HG100	39608381
12	52.5	38.0796562405754	=D12/10^5*Fitting!HG110	37967474
13	57.5	78.8634562632419	=D13/10^5*Fitting!HG120	36305238
14	62.5	139.37967747673	=D14/10^5*Fitting!HG130	32957586
15	67.5	212.899857330169	=D15/10^5*Fitting!HG140	28355398
16	72.5	283.783362668485	=D16/10^5*Fitting!HG150	22605333
17	77.5	345.975063590891	=D17/10^5*Fitting!HG160	16280034
18	82.5	370.732443186415	=D18/10^5*Fitting!HG170	10309444
19	87.5	371.514165627168	=D19/10^5*Fitting!HG180	5368021
20	92.5	329.360330623958	=D20/10^5*Fitting!HG190	2099980
21	97.5	264.355395715704	=D21/10^5*Fitting!HG200	253614
22	102.5		=D22/10^5*Fitting!HG210	

Raw Data | Fitting

Fig. 43: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Parameter section)

(Note that for the same parameters used in the Mathematica™ and Matlab™ programs, this template calculates similar expected mortality rates).

Upper left section:

Columns A and C contain the name tags of the parameters used in the program. There are two other parameters:

Cell B4 = $2\tau^2 r_i r_j N_{max} \alpha \log(2) r_A (\alpha_C - \beta_C)/\alpha_C$

Cell D8 = det = $-r_A (\alpha/(\alpha-\beta))^2 (\alpha_C - \beta_C)/\alpha_C$

used to facilitate the template, since these two parameters are used in multiple cells. The first parameter is just the product of the constant term K_h and the constant terms after one takes the derivative of the promotion term. The second parameter is just the constants inside the exponential of the promotion term (a and t are variables):

$$\textbf{Derivative of the promotion term}\left(1 - E^{-r_A \left(\frac{\alpha}{\alpha-\beta}\right)^2 \left(2^{(\alpha-\beta)(t-a)}-1\right) \frac{\alpha_C-\beta_C}{\alpha_C}}\right) = \frac{2^{(-a+t)(\alpha-\beta)}\, e^{\frac{\left(-1+2^{(-a+t)(\alpha-\beta)}\right)\alpha^2 r_A (\alpha_C-\beta_C)}{(\alpha-\beta)^2 \alpha_C}}\, \alpha^2 \text{Log}[2]\, r_A\, (\alpha_C - \beta_C)}{(\alpha - \beta)\, \alpha_C}$$

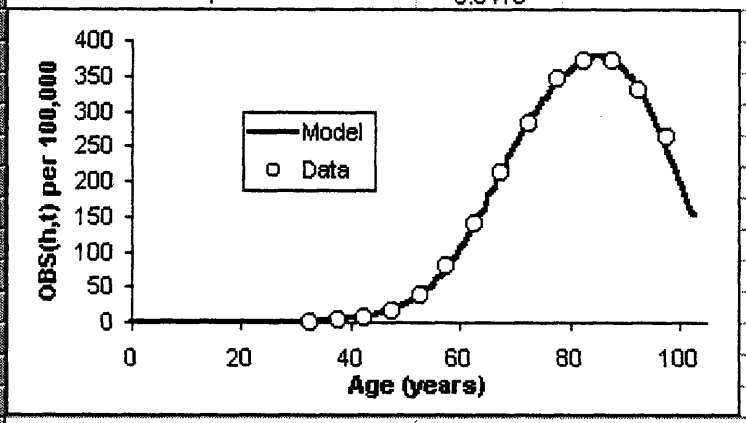
Lower left section:

Contains the calculated errors for each age point, and the sum of the errors.

Microsoft Excel - N2M1

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E
1	N_{max}	2.40E+08			
2	τ	5.7	$ri * 10^5 = rj / 3$	3.90E-02	
3	$(\alpha_C - \beta_C)/\alpha_C$	0.3365	Calculated Δ	48.17	
4	$2\tau^2 r_i r_j N_{max} \alpha \log(2) r_A (\alpha_C - \beta_C)/\alpha_C$	3.27E-09	ζ	1.17	Age at Initiation (a)→ Age at death (t) ↓
5	α	12.7	η	0.159	0
6	$r_A * 10^5$	1.55E-02	stopgrowth	16.5	0.5
7	$\alpha - \beta = \mu$	0.2434	$N_{1.5}$	45946740	1
8	F	0.9617	$-r_A (\alpha/(\alpha-\beta))^2 (\alpha_C - \beta_C)/\alpha_C$	-1.42E-04	1.5
9	f	0.0418			2
10					2.5
11					3
12					3.5
13					4
14					4.5
15					5
16					5.5
17					6
18					6.5
19					7
20					7.5
21					8
22					8.5
23		Error at age:	Sum of errors		9
24	#VALUE!	3	13417.17662		9.5
25	#VALUE!	7.5			10
26	#VALUE!	12.5			10.5
27	#VALUE!	17.5			11
28	#VALUE!	22.5			11.5
29	#VALUE!	27.5			12
30	24.56170544	32.5			12.5
31	352.8963361	37.5			13
32	1295.440028	42.5			13.5
33	1054.183287	47.5			14
34	99.65488053	52.5			14.5
35	956.9150598	57.5			15
36	800.909421	62.5			15.5
37	419.5035041	67.5			16
38	2876.289137	72.5			16.5
39	4507.014417	77.5			17
40	138.2997616	82.5			17.5
41	121.8531265	87.5			18
42	69.1845084	92.5			18.5
43	700.4714443	97.5			19
44	#VALUE!	102.5			19.5



Raw Data | Fitting

Fig. 44: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Parameter section) w/ formulae

Instructions:

The following cells must be filled in,

Cell B1: Number of cells in adult, N_{max}

Cell B2: Stochastic survival probability of cancerous cell, $(\alpha_c - \beta_c)/\alpha_c$

Cell B3: Normal cell turnover rate, τ

Cell B5: Precancerous cell division rate, α

Cell B6: Promotion mutation rate, r_A (multiplied by 100,000; Excel™'s maximum likelihood routine does not work well when changing cells with small values. r_A is one of the terms that the program will try to alter to maximize fit, so it has been multiplied factor of 100,000. Note that any reference to this cell must take this factor into account)

Cell B7: Growth rate of precancerous lesion, α - β

Cell B8: Fraction at risk, F

Cell B9: Correction factor for connected forms of death, f

Cell D2: First initiation mutation rare r_i; here assumed to be a third of the rate of the second initiation mutation rate, r_j (multiplied by 100,000; see explanation for Cell B6)

Cell D4: Growth rate of child age 0 to 1.5, ζ

Cell D5: Growth rate of child age 1.5 through puberty, η

Cell D6: Age at which child stops growing

All other cells contain formulae written in terms of the above defined cells. Excel™ formulae can refer to a cell by its column and row (letter, number); instead, we opted to use a feature of Excel™ that gives cells actual names. To do so, select cell (click on it), go to the Insert Menu, click on Name:Define. This gives you the option to give the selected cell a name.

For example Cell D8:

```
=-rA/100000*(alpha/mu)^2*survc
```

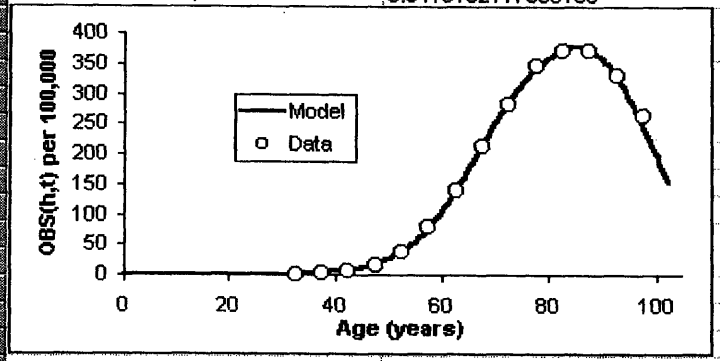
refers to rA (Cell B6 divided back by 100,000 to get true value), alpha (Cell B5), mu (Cell B7), and survc (Cell B2).

The error cells are references to the HJ column which is explained in Figures 49 and 51

Microsoft Excel - N2M1 Formulas

File Edit View Insert Format Tools Data Window Help

	A	B	C	D
1	N_{max}	240000000		
2	τ	5.7	$ri * 10^5 = rj / 3$	0.0390054960031954
3	$(\alpha_C - \beta_C)/\alpha_C$	=(42.2-28)/42.2	Calculated Δ	=LOG(1+(rA/100000*(alpha/(mu*LN(2)))^2*survc)^-1,2)/mu
4	$2\tau^2 r_i r_j N_{max} \alpha \log(2) r_A (\alpha_C - \beta_C)/\alpha_C$	=2*tau^2*ri^2*3*Nmax*alpha*LN(2)*rA/(100000)^3*survc	ζ	1.17
5	α	12.7	η	0.159
6	$r_A * 10^5$	0.0155216772589471	stopgrowth	16.5
7	$\alpha - \beta = \mu$	0.243355294945299	$N_{1.5}$	=Nmax/(2^(eta*(stopgrowth-1.5)))
8	F	0.961726478844813	$-r_A (\alpha/(\alpha-\beta))^2 (\alpha_C - \beta_C)/\alpha_C$	=-rA/100000*(alpha/mu)^2*survc
9	f	0.0418162117835156		
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23		Error at age:	Sum of errors	
24	=HJ11	3	=SUMIF(A24:A44,">0")	
25	=HJ20	7.5		
26	=HJ30	12.5		
27	=HJ40	17.5		
28	=HJ50	22.5		
29	=HJ60	27.5		
30	=HJ70	32.5		
31	=HJ80	37.5		
32	=HJ90	42.5		
33	=HJ100	47.5		
34	=HJ110	52.5		
35	=HJ120	57.5		
36	=HJ130	62.5		
37	=HJ140	67.5		



Fitting

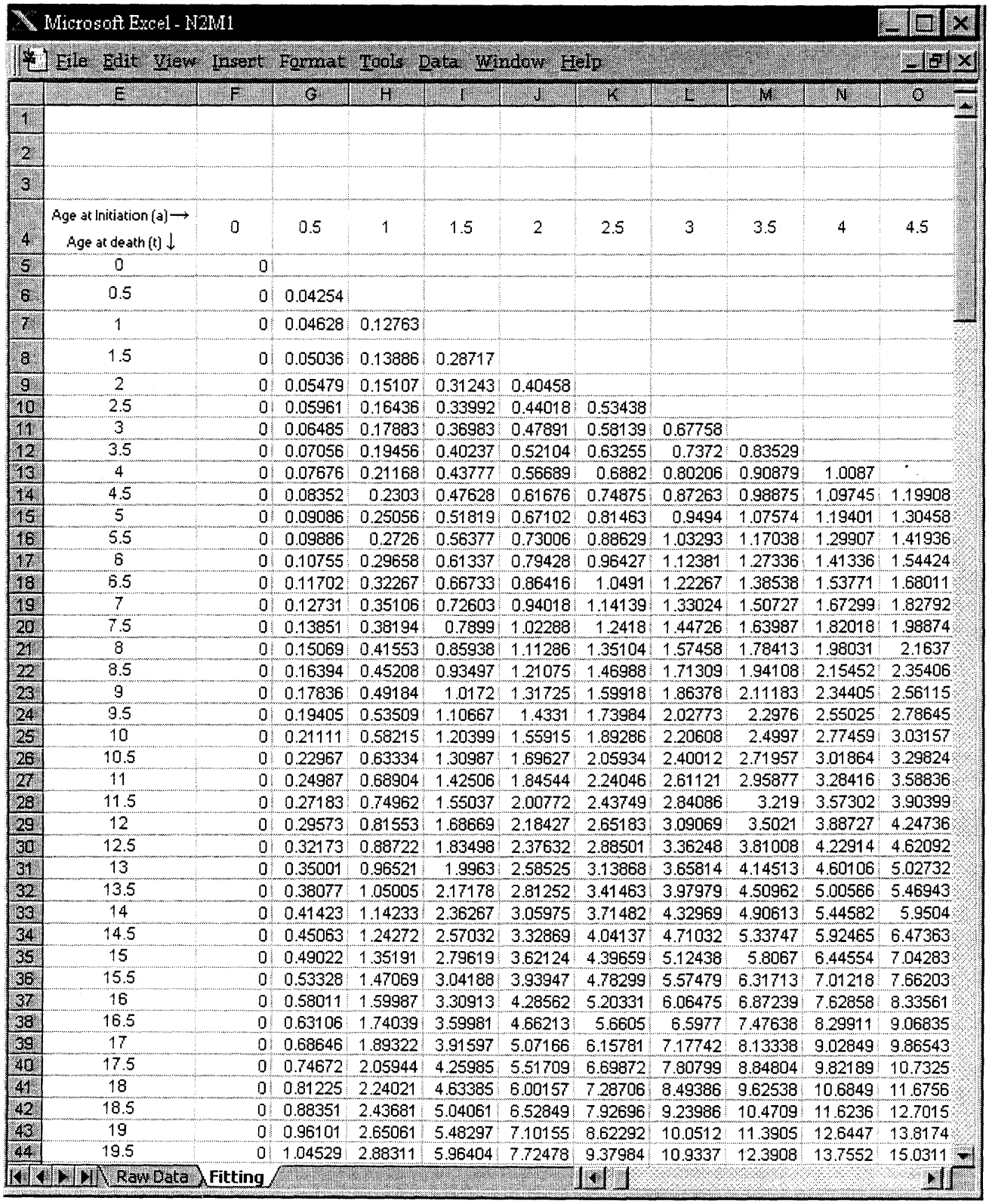
Fig. 45: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Contribution to death at age t, from lesion initiated at age a)

Due to limited space only

shown for ages t (0,0.5,1,1.5,....4.5)

shown for ages a (0,0.5,1,1.5,....19.5),

but are actually calculated up to age 102.5. Note that we do not calculate for a > t, since an individual cannot die at age t, from a precancerous lesion that was initiated at a later age.



	E	F	G	H	I	J	K	L	M	N	O
1											
2											
3											
4	Age at Initiation (a)→ Age at death (t) ↓	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5
5	0	0									
6	0.5	0	0.04254								
7	1	0	0.04628	0.12763							
8	1.5	0	0.05036	0.13886	0.28717						
9	2	0	0.05479	0.15107	0.31243	0.40458					
10	2.5	0	0.05961	0.16436	0.33992	0.44018	0.53438				
11	3	0	0.06485	0.17883	0.36983	0.47891	0.58139	0.67758			
12	3.5	0	0.07056	0.19456	0.40237	0.52104	0.63255	0.7372	0.83529		
13	4	0	0.07676	0.21168	0.43777	0.56689	0.6882	0.80206	0.90879	1.0087	
14	4.5	0	0.08352	0.2303	0.47628	0.61676	0.74875	0.87263	0.98875	1.09745	1.19908
15	5	0	0.09086	0.25056	0.51819	0.67102	0.81463	0.9494	1.07574	1.19401	1.30458
16	5.5	0	0.09886	0.2726	0.56377	0.73006	0.88629	1.03293	1.17038	1.29907	1.41936
17	6	0	0.10755	0.29658	0.61337	0.79428	0.96427	1.12381	1.27336	1.41336	1.54424
18	6.5	0	0.11702	0.32267	0.66733	0.86416	1.0491	1.22267	1.38538	1.53771	1.68011
19	7	0	0.12731	0.35106	0.72603	0.94018	1.14139	1.33024	1.50727	1.67299	1.82792
20	7.5	0	0.13851	0.38194	0.7899	1.02288	1.2418	1.44726	1.63987	1.82018	1.98874
21	8	0	0.15069	0.41553	0.85938	1.11286	1.35104	1.57458	1.78413	1.98031	2.1637
22	8.5	0	0.16394	0.45208	0.93497	1.21075	1.46988	1.71309	1.94108	2.15452	2.35406
23	9	0	0.17836	0.49184	1.0172	1.31725	1.59918	1.86378	2.11183	2.34405	2.56115
24	9.5	0	0.19405	0.53509	1.10667	1.4331	1.73984	2.02773	2.2976	2.55025	2.78645
25	10	0	0.21111	0.58215	1.20399	1.55915	1.89286	2.20608	2.4997	2.77459	3.03157
26	10.5	0	0.22967	0.63334	1.30987	1.69627	2.05934	2.40012	2.71957	3.01864	3.29824
27	11	0	0.24987	0.68904	1.42506	1.84544	2.24046	2.61121	2.95877	3.28416	3.58836
28	11.5	0	0.27183	0.74962	1.55037	2.00772	2.43749	2.84086	3.219	3.57302	3.90399
29	12	0	0.29573	0.81553	1.68669	2.18427	2.65183	3.09069	3.5021	3.88727	4.24736
30	12.5	0	0.32173	0.88722	1.83498	2.37632	2.88501	3.36248	3.81008	4.22914	4.62092
31	13	0	0.35001	0.96521	1.9963	2.58525	3.13868	3.65814	4.14513	4.60106	5.02732
32	13.5	0	0.38077	1.05005	2.17178	2.81252	3.41463	3.97979	4.50962	5.00566	5.46943
33	14	0	0.41423	1.14233	2.36267	3.05975	3.71482	4.32969	4.90613	5.44582	5.9504
34	14.5	0	0.45063	1.24272	2.57032	3.32869	4.04137	4.71032	5.33747	5.92465	6.47363
35	15	0	0.49022	1.35191	2.79619	3.62124	4.39659	5.12438	5.8067	6.44554	7.04283
36	15.5	0	0.53328	1.47069	3.04188	3.93947	4.78299	5.57479	6.31713	7.01218	7.66203
37	16	0	0.58011	1.59987	3.30913	4.28562	5.20331	6.06475	6.87239	7.62858	8.33561
38	16.5	0	0.63106	1.74039	3.59981	4.66213	5.6605	6.5977	7.47638	8.29911	9.06835
39	17	0	0.68646	1.89322	3.91597	5.07166	6.15781	7.17742	8.13338	9.02849	9.86543
40	17.5	0	0.74672	2.05944	4.25985	5.51709	6.69872	7.80799	8.84804	9.82189	10.7325
41	18	0	0.81225	2.24021	4.63385	6.00157	7.28706	8.49386	9.62538	10.6849	11.6756
42	18.5	0	0.88351	2.43681	5.04061	6.52849	7.92696	9.23986	10.4709	11.6236	12.7015
43	19	0	0.96101	2.65061	5.48297	7.10155	8.62292	10.0512	11.3905	12.6447	13.8174
44	19.5	0	1.04529	2.88311	5.96404	7.72478	9.37984	10.9337	12.3908	13.7552	15.0311

Fig. 46: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Contribution to death at age t, from lesion initiated at age a) w/ formulae (a <= 1.5)

The formula here consist only of the terms inside the integration term of $P_{OBS}(h,t)$. We have not yet multiplied the constant of cell B4 (See Figure 43 legend for explanation of this constant) because every cell must be multiplied by this term. Rather than use up memory to do so, we wait until integration of these cells to do so. (Figure 49). The term det is defined as Cell D8 (See Figure 43 legend for explanation of this constant).

The G column shows the formulae for the lesions initiated at age 0.5. Reference G4 is for the age 'a' at initiation; unlike other cells, we did not name G4 as this is a variable.

References E6, E7, E8, ... are for the age 't' at death. Likewise, we cannot name them as these are variables.

(See the Mathematica™ for confirmation that these are indeed the correct formulas, Section 3.6.1)

Microsoft Excel - N2M1 Formulas

File Edit View Insert Format Tools Data Window Help

	E	F	G
1			
2			
3			
4	Age at Initiation (a)→ Age at death (t) ↓	0	0.5
5	0	0	
6	0.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E6-G4)*mu)*EXP((2^((E6-G4)*mu)-1)*det)
7	1	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E7-G4)*mu)*EXP((2^((E7-G4)*mu)-1)*det)
8	1.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E8-G4)*mu)*EXP((2^((E8-G4)*mu)-1)*det)
9	2	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E9-G4)*mu)*EXP((2^((E9-G4)*mu)-1)*det)
10	2.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E10-G4)*mu)*EXP((2^((E10-G4)*mu)-1)*det)
11	3	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E11-G4)*mu)*EXP((2^((E11-G4)*mu)-1)*det)
12	3.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E12-G4)*mu)*EXP((2^((E12-G4)*mu)-1)*det)
13	4	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E13-G4)*mu)*EXP((2^((E13-G4)*mu)-1)*det)
14	4.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E14-G4)*mu)*EXP((2^((E14-G4)*mu)-1)*det)
15	5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E15-G4)*mu)*EXP((2^((E15-G4)*mu)-1)*det)
16	5.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E16-G4)*mu)*EXP((2^((E16-G4)*mu)-1)*det)
17	6	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E17-G4)*mu)*EXP((2^((E17-G4)*mu)-1)*det)
18	6.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E18-G4)*mu)*EXP((2^((E18-G4)*mu)-1)*det)
19	7	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E19-G4)*mu)*EXP((2^((E19-G4)*mu)-1)*det)
20	7.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E20-G4)*mu)*EXP((2^((E20-G4)*mu)-1)*det)
21	8	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E21-G4)*mu)*EXP((2^((E21-G4)*mu)-1)*det)
22	8.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E22-G4)*mu)*EXP((2^((E22-G4)*mu)-1)*det)
23	9	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E23-G4)*mu)*EXP((2^((E23-G4)*mu)-1)*det)
24	9.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E24-G4)*mu)*EXP((2^((E24-G4)*mu)-1)*det)
25	10	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E25-G4)*mu)*EXP((2^((E25-G4)*mu)-1)*det)
26	10.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E26-G4)*mu)*EXP((2^((E26-G4)*mu)-1)*det)
27	11	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E27-G4)*mu)*EXP((2^((E27-G4)*mu)-1)*det)
28	11.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E28-G4)*mu)*EXP((2^((E28-G4)*mu)-1)*det)
29	12	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E29-G4)*mu)*EXP((2^((E29-G4)*mu)-1)*det)
30	12.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E30-G4)*mu)*EXP((2^((E30-G4)*mu)-1)*det)
31	13	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E31-G4)*mu)*EXP((2^((E31-G4)*mu)-1)*det)
32	13.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E32-G4)*mu)*EXP((2^((E32-G4)*mu)-1)*det)
33	14	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E33-G4)*mu)*EXP((2^((E33-G4)*mu)-1)*det)
34	14.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E34-G4)*mu)*EXP((2^((E34-G4)*mu)-1)*det)
35	15	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E35-G4)*mu)*EXP((2^((E35-G4)*mu)-1)*det)
36	15.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E36-G4)*mu)*EXP((2^((E36-G4)*mu)-1)*det)
37	16	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E37-G4)*mu)*EXP((2^((E37-G4)*mu)-1)*det)
38	16.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E38-G4)*mu)*EXP((2^((E38-G4)*mu)-1)*det)
39	17	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E39-G4)*mu)*EXP((2^((E39-G4)*mu)-1)*det)
40	17.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E40-G4)*mu)*EXP((2^((E40-G4)*mu)-1)*det)
41	18	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E41-G4)*mu)*EXP((2^((E41-G4)*mu)-1)*det)
42	18.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E42-G4)*mu)*EXP((2^((E42-G4)*mu)-1)*det)
43	19	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E43-G4)*mu)*EXP((2^((E43-G4)*mu)-1)*det)
44	19.5	0	=G4*N1.5/Nmax*2^(zeta*(G4-1.5))*2^((E44-G4)*mu)*EXP((2^((E44-G4)*mu)-1)*det)

Raw Data | Fitting

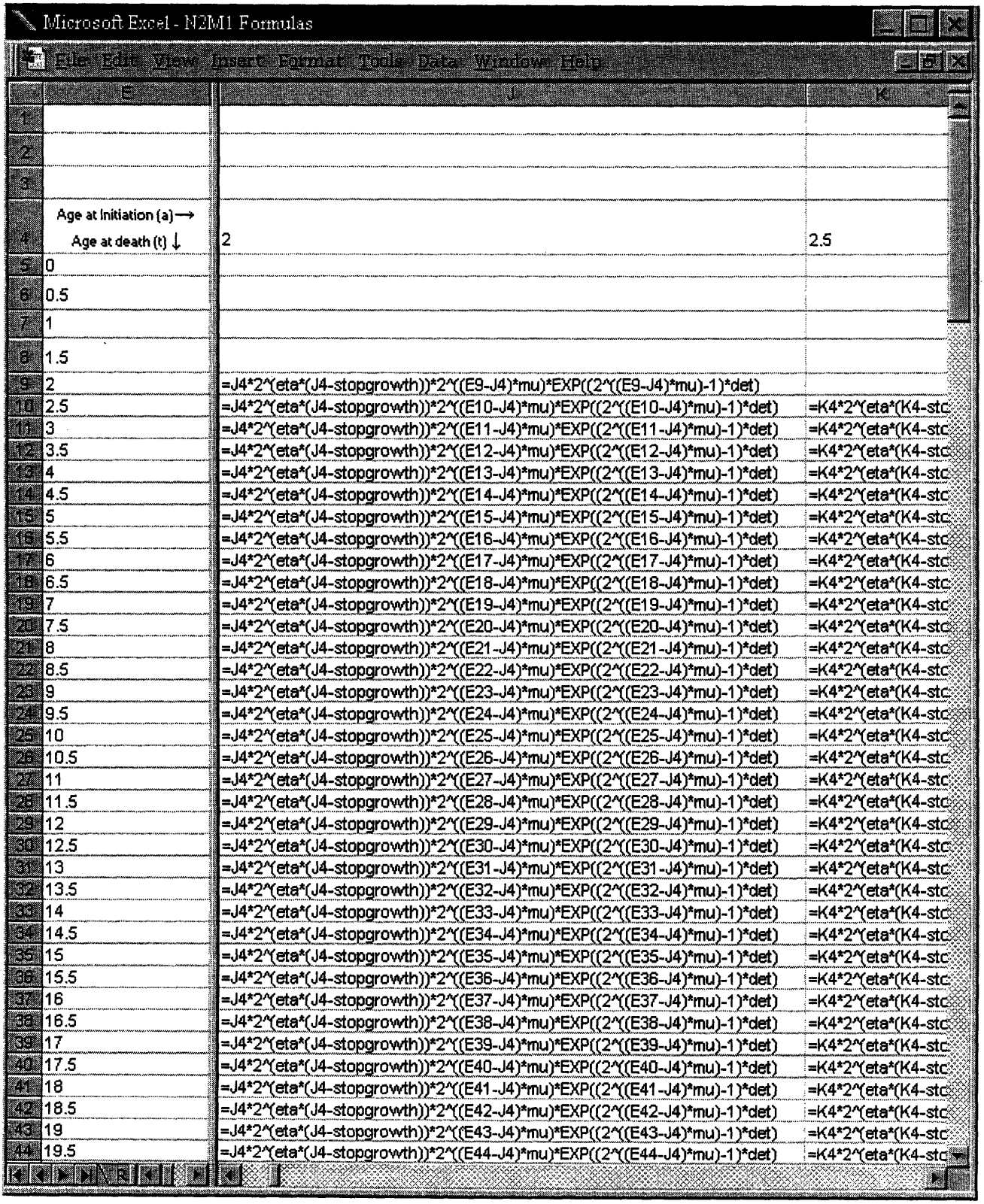
Fig. 47: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Contribution to death at age t, from lesion initiated at age a) w/ formulae (1.5 < a <= adulthood)

The formula here consist again only of the terms inside the integration term of $P_{OBS}(h,t)$. We have not yet multiplied the constant from Cell B4 (Figure 43), because every cell must be multiplied by this term. Rather than use up memory to do so, we wait until integration of these cells to do so. (Figure 49). The term det is defined as Cell D8 (Figure 43).

The J column shows the formulae for the lesions initiated at age 2. Reference J4 is for the age 'a' at initiation; unlike other cells, we did not name J4 as this is a variable.

References E6, E7, E8, ... are for the age 't' at death. Likewise, we cannot name them as these are variables.

(See the Mathematica™ for confirmation that these are indeed the correct formulas, Section 3.6.1)



Microsoft Excel - N2M1 Formulas

File Edit View Insert Format Tools Data Window Help

	E	J	K
1			
2			
3			
4	Age at Initiation (a) → Age at death (t) ↓	2	2.5
5	0		
6	0.5		
7	1		
8	1.5		
9	2	=J4*2^(eta*(J4-stopgrowth))*2^((E9-J4)*mu)*EXP((2^((E9-J4)*mu)-1)*det)	
10	2.5	=J4*2^(eta*(J4-stopgrowth))*2^((E10-J4)*mu)*EXP((2^((E10-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
11	3	=J4*2^(eta*(J4-stopgrowth))*2^((E11-J4)*mu)*EXP((2^((E11-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
12	3.5	=J4*2^(eta*(J4-stopgrowth))*2^((E12-J4)*mu)*EXP((2^((E12-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
13	4	=J4*2^(eta*(J4-stopgrowth))*2^((E13-J4)*mu)*EXP((2^((E13-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
14	4.5	=J4*2^(eta*(J4-stopgrowth))*2^((E14-J4)*mu)*EXP((2^((E14-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
15	5	=J4*2^(eta*(J4-stopgrowth))*2^((E15-J4)*mu)*EXP((2^((E15-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
16	5.5	=J4*2^(eta*(J4-stopgrowth))*2^((E16-J4)*mu)*EXP((2^((E16-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
17	6	=J4*2^(eta*(J4-stopgrowth))*2^((E17-J4)*mu)*EXP((2^((E17-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
18	6.5	=J4*2^(eta*(J4-stopgrowth))*2^((E18-J4)*mu)*EXP((2^((E18-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
19	7	=J4*2^(eta*(J4-stopgrowth))*2^((E19-J4)*mu)*EXP((2^((E19-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
20	7.5	=J4*2^(eta*(J4-stopgrowth))*2^((E20-J4)*mu)*EXP((2^((E20-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
21	8	=J4*2^(eta*(J4-stopgrowth))*2^((E21-J4)*mu)*EXP((2^((E21-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
22	8.5	=J4*2^(eta*(J4-stopgrowth))*2^((E22-J4)*mu)*EXP((2^((E22-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
23	9	=J4*2^(eta*(J4-stopgrowth))*2^((E23-J4)*mu)*EXP((2^((E23-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
24	9.5	=J4*2^(eta*(J4-stopgrowth))*2^((E24-J4)*mu)*EXP((2^((E24-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
25	10	=J4*2^(eta*(J4-stopgrowth))*2^((E25-J4)*mu)*EXP((2^((E25-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
26	10.5	=J4*2^(eta*(J4-stopgrowth))*2^((E26-J4)*mu)*EXP((2^((E26-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
27	11	=J4*2^(eta*(J4-stopgrowth))*2^((E27-J4)*mu)*EXP((2^((E27-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
28	11.5	=J4*2^(eta*(J4-stopgrowth))*2^((E28-J4)*mu)*EXP((2^((E28-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
29	12	=J4*2^(eta*(J4-stopgrowth))*2^((E29-J4)*mu)*EXP((2^((E29-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
30	12.5	=J4*2^(eta*(J4-stopgrowth))*2^((E30-J4)*mu)*EXP((2^((E30-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
31	13	=J4*2^(eta*(J4-stopgrowth))*2^((E31-J4)*mu)*EXP((2^((E31-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
32	13.5	=J4*2^(eta*(J4-stopgrowth))*2^((E32-J4)*mu)*EXP((2^((E32-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
33	14	=J4*2^(eta*(J4-stopgrowth))*2^((E33-J4)*mu)*EXP((2^((E33-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
34	14.5	=J4*2^(eta*(J4-stopgrowth))*2^((E34-J4)*mu)*EXP((2^((E34-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
35	15	=J4*2^(eta*(J4-stopgrowth))*2^((E35-J4)*mu)*EXP((2^((E35-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
36	15.5	=J4*2^(eta*(J4-stopgrowth))*2^((E36-J4)*mu)*EXP((2^((E36-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
37	16	=J4*2^(eta*(J4-stopgrowth))*2^((E37-J4)*mu)*EXP((2^((E37-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
38	16.5	=J4*2^(eta*(J4-stopgrowth))*2^((E38-J4)*mu)*EXP((2^((E38-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
39	17	=J4*2^(eta*(J4-stopgrowth))*2^((E39-J4)*mu)*EXP((2^((E39-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
40	17.5	=J4*2^(eta*(J4-stopgrowth))*2^((E40-J4)*mu)*EXP((2^((E40-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
41	18	=J4*2^(eta*(J4-stopgrowth))*2^((E41-J4)*mu)*EXP((2^((E41-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
42	18.5	=J4*2^(eta*(J4-stopgrowth))*2^((E42-J4)*mu)*EXP((2^((E42-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
43	19	=J4*2^(eta*(J4-stopgrowth))*2^((E43-J4)*mu)*EXP((2^((E43-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc
44	19.5	=J4*2^(eta*(J4-stopgrowth))*2^((E44-J4)*mu)*EXP((2^((E44-J4)*mu)-1)*det)	=K4*2^(eta*(K4-stc

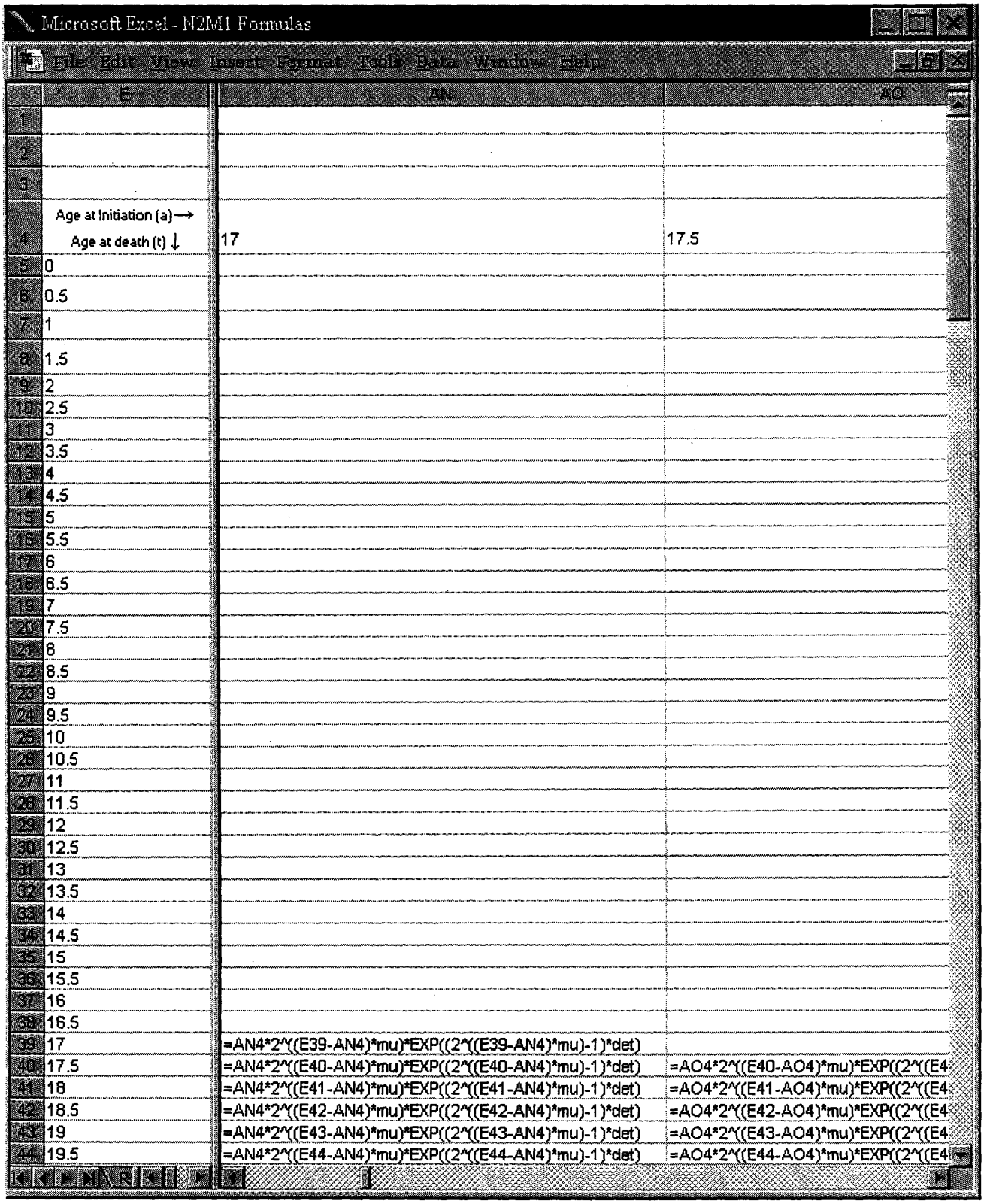
Fig. 48: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Contribution to death at age t, from lesion initiated at age a) w/ formulae (a > adulthood)

The formula here consist again only of the terms inside the integration term of $P_{OBS}(h,t)$. We have not yet multiplied the constant in Cell B4 (Figure 43), because every cell must be multiplied by this term. Rather than use up memory to do so, we wait until integration of these cells to do so. (Figure 49). The term det is defined as Cell D8 (Figure 43).

The AN column shows the formulas for the lesions initiated at age 17. Reference AN4 is for the age 'a' at initiation; unlike other cells, we did not name AN4 as this is a variable.

References E6, E7, E8, ... are for the age 't' at death. Likewise, we cannot name them as these are variables.

(See the Mathematica™ for confirmation that these are indeed the correct formulas, Section 3.6.1)


Microsoft Excel - N2M1 Formulas

File Edit View Insert Format Tools Data Window Help

	E	AN	AO
1			
2			
3			
4	Age at Initiation (a)→ Age at death (t) ↓	17	17.5
5	0		
6	0.5		
7	1		
8	1.5		
9	2		
10	2.5		
11	3		
12	3.5		
13	4		
14	4.5		
15	5		
16	5.5		
17	6		
18	6.5		
19	7		
20	7.5		
21	8		
22	8.5		
23	9		
24	9.5		
25	10		
26	10.5		
27	11		
28	11.5		
29	12		
30	12.5		
31	13		
32	13.5		
33	14		
34	14.5		
35	15		
36	15.5		
37	16		
38	16.5		
39	17	=AN4*2^((E39-AN4)*mu)*EXP((2^((E39-AN4)*mu)-1)*det)	
40	17.5	=AN4*2^((E40-AN4)*mu)*EXP((2^((E40-AN4)*mu)-1)*det)	=AO4*2^((E40-AO4)*mu)*EXP((2^((E4
41	18	=AN4*2^((E41-AN4)*mu)*EXP((2^((E41-AN4)*mu)-1)*det)	=AO4*2^((E41-AO4)*mu)*EXP((2^((E4
42	18.5	=AN4*2^((E42-AN4)*mu)*EXP((2^((E42-AN4)*mu)-1)*det)	=AO4*2^((E42-AO4)*mu)*EXP((2^((E4
43	19	=AN4*2^((E43-AN4)*mu)*EXP((2^((E43-AN4)*mu)-1)*det)	=AO4*2^((E43-AO4)*mu)*EXP((2^((E4
44	19.5	=AN4*2^((E44-AN4)*mu)*EXP((2^((E44-AN4)*mu)-1)*det)	=AO4*2^((E44-AO4)*mu)*EXP((2^((E4

Fig. 49: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Calculation of OBS(h,t))

Instructions:

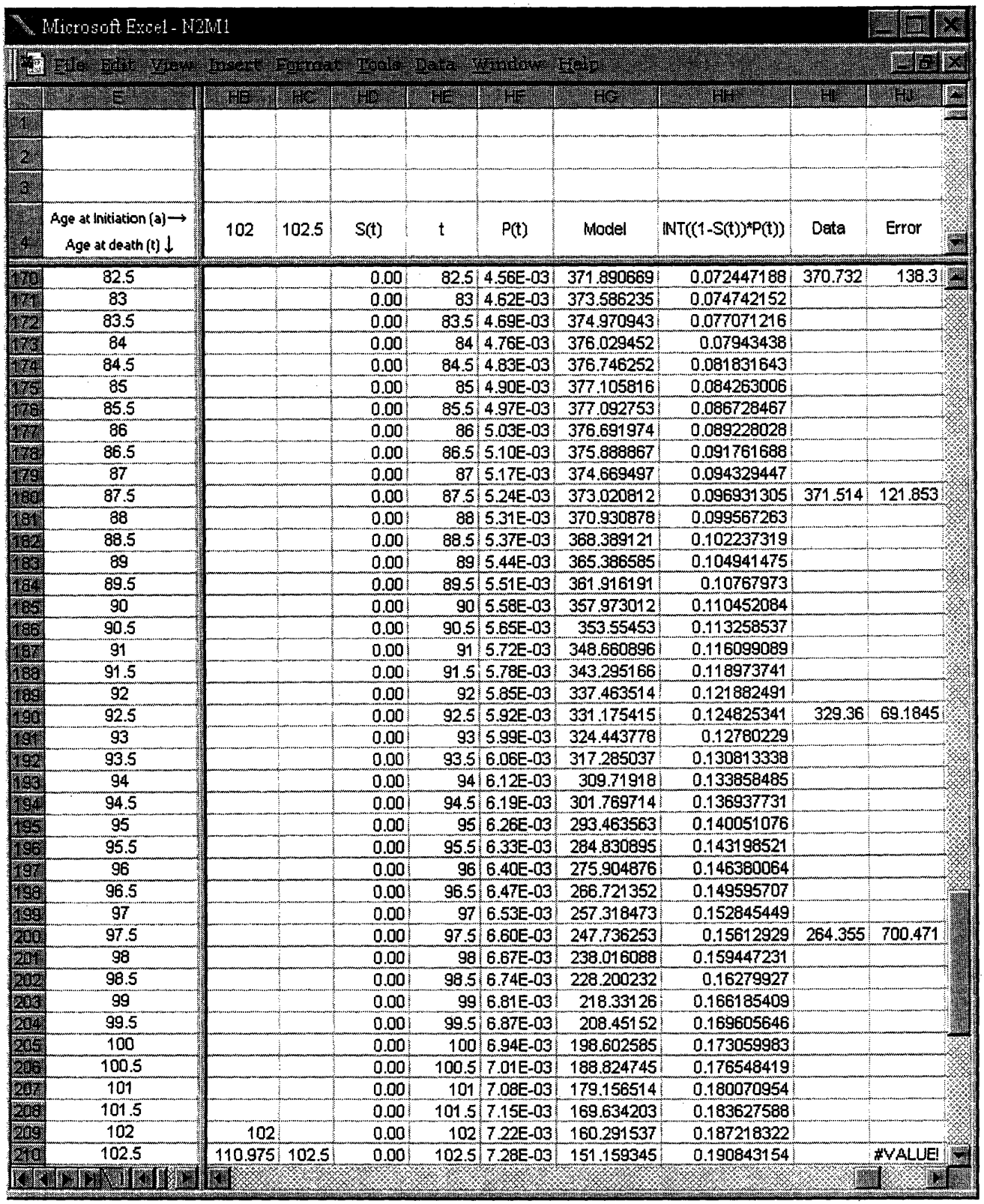
This part of the template contains the survival rates as a function of age which need to be typed into Column HD. All other cells are formulas.

Calculated cells:

The template calculates $P_{OBS}(h,t)$ (Column HF),
integrates the product of $P_{OBS}(h,t)$ $(1 - S(h,t))$ (Column HH)
and then calculates OBS(h,t) (Column HG)

Column HI just refers back to the Raw Data worksheet to acquire the observed data.
Column HJ then calculates the error (given a weighing factor), comparing the model (OBS(h,t)) to the raw data.

The plot of the observed and calculated mortality data of Figure 43 uses Columns HG, HI



Microsoft Excel - N2M1

File Edit View Insert Format Tools Data Window Help

	E	HB	HC	HD	HE	HF	HG	HH	HI	HJ
1										
2										
3										
4	Age at Initiation (a)→ Age at death (t) ↓	102	102.5	S(t)	t	P(t)	Model	INT((1-S(t))*P(t))	Data	Error
170	82.5			0.00	82.5	4.56E-03	371.890669	0.072447188	370.732	138.3
171	83			0.00	83	4.62E-03	373.586235	0.074742152		
172	83.5			0.00	83.5	4.69E-03	374.970943	0.077071216		
173	84			0.00	84	4.76E-03	376.029452	0.07943438		
174	84.5			0.00	84.5	4.83E-03	376.746252	0.081831643		
175	85			0.00	85	4.90E-03	377.105816	0.084263006		
176	85.5			0.00	85.5	4.97E-03	377.092753	0.086728467		
177	86			0.00	86	5.03E-03	376.691974	0.089228028		
178	86.5			0.00	86.5	5.10E-03	375.888867	0.091761688		
179	87			0.00	87	5.17E-03	374.669497	0.094329447		
180	87.5			0.00	87.5	5.24E-03	373.020812	0.096931305	371.514	121.853
181	88			0.00	88	5.31E-03	370.930878	0.099567263		
182	88.5			0.00	88.5	5.37E-03	368.389121	0.102237319		
183	89			0.00	89	5.44E-03	365.386585	0.104941475		
184	89.5			0.00	89.5	5.51E-03	361.916191	0.10767973		
185	90			0.00	90	5.58E-03	357.973012	0.110452084		
186	90.5			0.00	90.5	5.65E-03	353.55453	0.113258537		
187	91			0.00	91	5.72E-03	348.660896	0.116099089		
188	91.5			0.00	91.5	5.78E-03	343.295166	0.118973741		
189	92			0.00	92	5.85E-03	337.463514	0.121882491		
190	92.5			0.00	92.5	5.92E-03	331.175415	0.124825341	329.36	69.1845
191	93			0.00	93	5.99E-03	324.443778	0.12780229		
192	93.5			0.00	93.5	6.06E-03	317.285037	0.130813338		
193	94			0.00	94	6.12E-03	309.71918	0.133858485		
194	94.5			0.00	94.5	6.19E-03	301.769714	0.136937731		
195	95			0.00	95	6.26E-03	293.463563	0.140051076		
196	95.5			0.00	95.5	6.33E-03	284.830895	0.143198521		
197	96			0.00	96	6.40E-03	275.904876	0.146380064		
198	96.5			0.00	96.5	6.47E-03	266.721352	0.149595707		
199	97			0.00	97	6.53E-03	257.318473	0.152845449		
200	97.5			0.00	97.5	6.60E-03	247.736253	0.15612929	264.355	700.471
201	98			0.00	98	6.67E-03	238.016088	0.159447231		
202	98.5			0.00	98.5	6.74E-03	228.200232	0.16279927		
203	99			0.00	99	6.81E-03	218.33126	0.166185409		
204	99.5			0.00	99.5	6.87E-03	208.45152	0.169605646		
205	100			0.00	100	6.94E-03	198.602585	0.173059983		
206	100.5			0.00	100.5	7.01E-03	188.824745	0.176548419		
207	101			0.00	101	7.08E-03	179.156514	0.180070954		
208	101.5			0.00	101.5	7.15E-03	169.634203	0.183627588		
209	102	102		0.00	102	7.22E-03	160.291537	0.187218322		
210	102.5	110.975	102.5	0.00	102.5	7.28E-03	151.159345	0.190843154		#VALUE!

Fig. 50: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Calculation of OBS(h,t)) w/ formulae

To calculate $P_{OBS}(h,t)$ (Column HF), we must account for every single contribution to death at age t by every lesion (since each column represents the age 'a' at initiation, we summed these terms and divided by 2 as an approximation to the integral of $P_{OBS}(h,t)$). We now multiply by the constant terms found in POBS(h,t) (Cell B4 as explained in Figure 43).

(We used SUMIF(cells, ">0") instead of the regular SUM(cells) command because certain sets of parameters have been found to give errors (i.e. division by 0) which cannot be summed; these errors typically occur when the maximum likelihood routine tries parameters which are very small.)

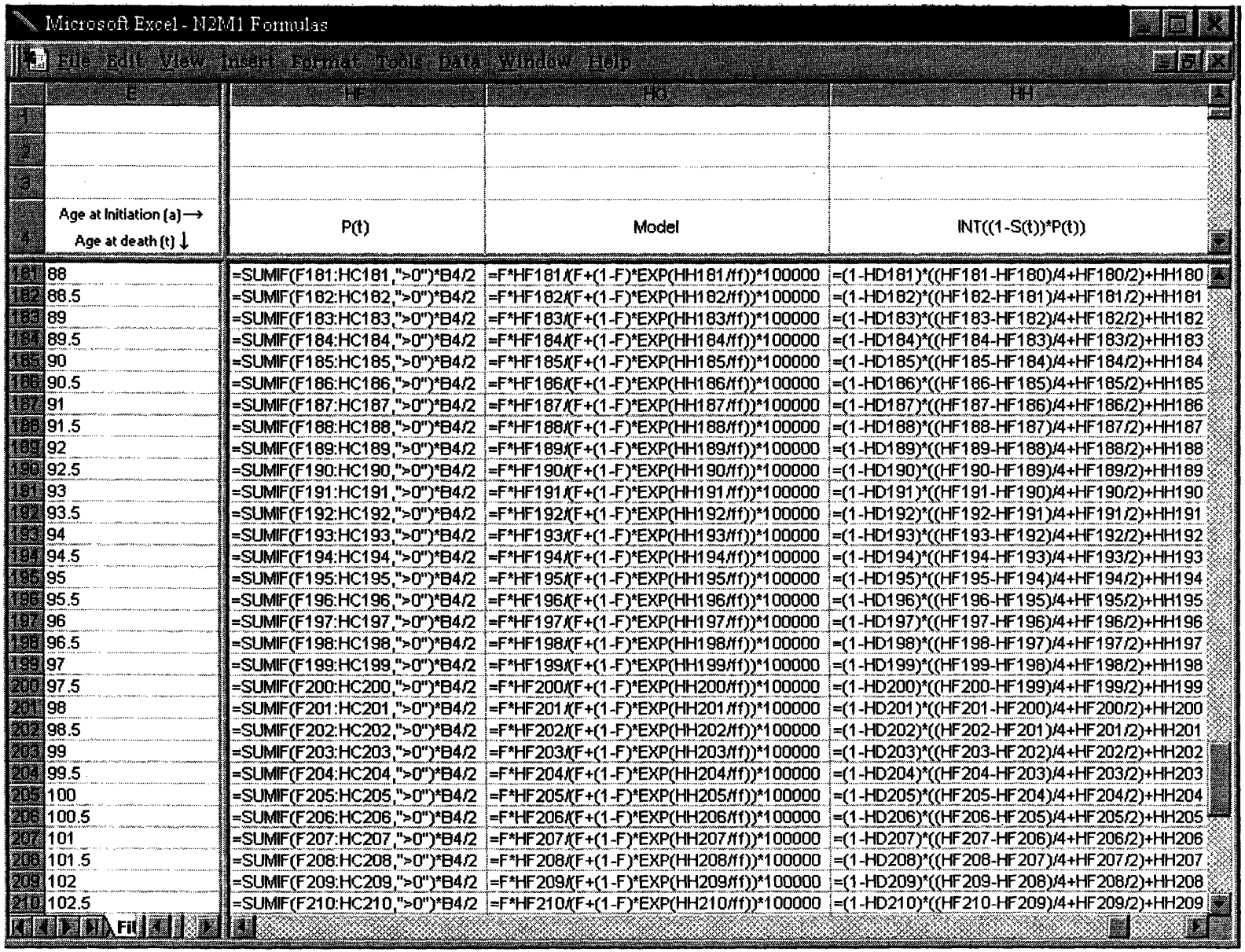
To calculate the integral of $P_{OBS}(h,t)$ $(1 - S(h,t))$ (Column HH), we used multiplication by trapezoids. For a monotonically increasing function, this is simply:

$$\sum_{i = \varepsilon,\ \text{intervals of } \varepsilon}^{t} [P_{OBS}(h,i)\,(1 - S(h,i)) + P_{OBS}(h,i - \varepsilon)\,(1 - S(i,t - \varepsilon))] \times \varepsilon / 2$$

The limit of the above equation as $\varepsilon \rightarrow 0$, is the exact integral. Since we have calculated initiation every 0.5 years, we use $\varepsilon = 0.5$. (The Matlab™ program had used $\varepsilon = 1$).

The calculation of OBS(h,t) (Column HG) is straightforward (Equation 16, Section 3.5.2.2). We multiply by 100,000 to express rates per 100,000.

(Excel™ is not case sensitive, so we used 'F' for the fraction at risk (Cell B8), and 'ff' for the correction factor for connected diseases (Cell B9))



Microsoft Excel - N2M1 Formulas

File Edit View Insert Format Tools Data Window Help

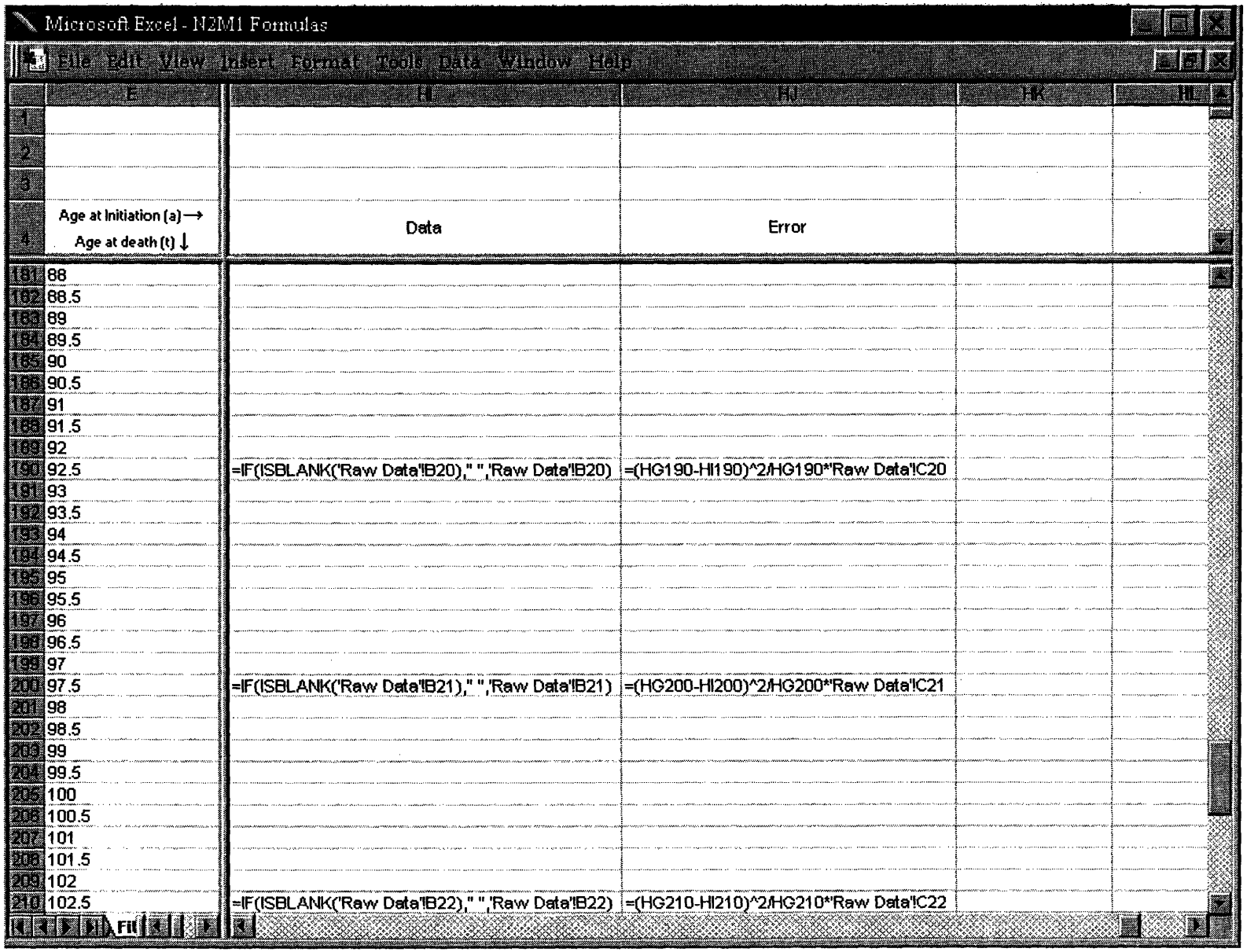
	E	HF	HG	HH
1				
2				
3				
4	Age at Initiation (a) → Age at death (t) ↓	P(t)	Model	INT((1-S(t))*P(t))
181	88	=SUMIF(F181:HC181,">0")*B4/2	=F*HF181/(F+(1-F)*EXP(HH181/ff))*100000	=(1-HD181)*((HF181-HF180)/4+HF180/2)+HH180
182	88.5	=SUMIF(F182:HC182,">0")*B4/2	=F*HF182/(F+(1-F)*EXP(HH182/ff))*100000	=(1-HD182)*((HF182-HF181)/4+HF181/2)+HH181
183	89	=SUMIF(F183:HC183,">0")*B4/2	=F*HF183/(F+(1-F)*EXP(HH183/ff))*100000	=(1-HD183)*((HF183-HF182)/4+HF182/2)+HH182
184	89.5	=SUMIF(F184:HC184,">0")*B4/2	=F*HF184/(F+(1-F)*EXP(HH184/ff))*100000	=(1-HD184)*((HF184-HF183)/4+HF183/2)+HH183
185	90	=SUMIF(F185:HC185,">0")*B4/2	=F*HF185/(F+(1-F)*EXP(HH185/ff))*100000	=(1-HD185)*((HF185-HF184)/4+HF184/2)+HH184
186	90.5	=SUMIF(F186:HC186,">0")*B4/2	=F*HF186/(F+(1-F)*EXP(HH186/ff))*100000	=(1-HD186)*((HF186-HF185)/4+HF185/2)+HH185
187	91	=SUMIF(F187:HC187,">0")*B4/2	=F*HF187/(F+(1-F)*EXP(HH187/ff))*100000	=(1-HD187)*((HF187-HF186)/4+HF186/2)+HH186
188	91.5	=SUMIF(F188:HC188,">0")*B4/2	=F*HF188/(F+(1-F)*EXP(HH188/ff))*100000	=(1-HD188)*((HF188-HF187)/4+HF187/2)+HH187
189	92	=SUMIF(F189:HC189,">0")*B4/2	=F*HF189/(F+(1-F)*EXP(HH189/ff))*100000	=(1-HD189)*((HF189-HF188)/4+HF188/2)+HH188
190	92.5	=SUMIF(F190:HC190,">0")*B4/2	=F*HF190/(F+(1-F)*EXP(HH190/ff))*100000	=(1-HD190)*((HF190-HF189)/4+HF189/2)+HH189
191	93	=SUMIF(F191:HC191,">0")*B4/2	=F*HF191/(F+(1-F)*EXP(HH191/ff))*100000	=(1-HD191)*((HF191-HF190)/4+HF190/2)+HH190
192	93.5	=SUMIF(F192:HC192,">0")*B4/2	=F*HF192/(F+(1-F)*EXP(HH192/ff))*100000	=(1-HD192)*((HF192-HF191)/4+HF191/2)+HH191
193	94	=SUMIF(F193:HC193,">0")*B4/2	=F*HF193/(F+(1-F)*EXP(HH193/ff))*100000	=(1-HD193)*((HF193-HF192)/4+HF192/2)+HH192
194	94.5	=SUMIF(F194:HC194,">0")*B4/2	=F*HF194/(F+(1-F)*EXP(HH194/ff))*100000	=(1-HD194)*((HF194-HF193)/4+HF193/2)+HH193
195	95	=SUMIF(F195:HC195,">0")*B4/2	=F*HF195/(F+(1-F)*EXP(HH195/ff))*100000	=(1-HD195)*((HF195-HF194)/4+HF194/2)+HH194
196	95.5	=SUMIF(F196:HC196,">0")*B4/2	=F*HF196/(F+(1-F)*EXP(HH196/ff))*100000	=(1-HD196)*((HF196-HF195)/4+HF195/2)+HH195
197	96	=SUMIF(F197:HC197,">0")*B4/2	=F*HF197/(F+(1-F)*EXP(HH197/ff))*100000	=(1-HD197)*((HF197-HF196)/4+HF196/2)+HH196
198	96.5	=SUMIF(F198:HC198,">0")*B4/2	=F*HF198/(F+(1-F)*EXP(HH198/ff))*100000	=(1-HD198)*((HF198-HF197)/4+HF197/2)+HH197
199	97	=SUMIF(F199:HC199,">0")*B4/2	=F*HF199/(F+(1-F)*EXP(HH199/ff))*100000	=(1-HD199)*((HF199-HF198)/4+HF198/2)+HH198
200	97.5	=SUMIF(F200:HC200,">0")*B4/2	=F*HF200/(F+(1-F)*EXP(HH200/ff))*100000	=(1-HD200)*((HF200-HF199)/4+HF199/2)+HH199
201	98	=SUMIF(F201:HC201,">0")*B4/2	=F*HF201/(F+(1-F)*EXP(HH201/ff))*100000	=(1-HD201)*((HF201-HF200)/4+HF200/2)+HH200
202	98.5	=SUMIF(F202:HC202,">0")*B4/2	=F*HF202/(F+(1-F)*EXP(HH202/ff))*100000	=(1-HD202)*((HF202-HF201)/4+HF201/2)+HH201
203	99	=SUMIF(F203:HC203,">0")*B4/2	=F*HF203/(F+(1-F)*EXP(HH203/ff))*100000	=(1-HD203)*((HF203-HF202)/4+HF202/2)+HH202
204	99.5	=SUMIF(F204:HC204,">0")*B4/2	=F*HF204/(F+(1-F)*EXP(HH204/ff))*100000	=(1-HD204)*((HF204-HF203)/4+HF203/2)+HH203
205	100	=SUMIF(F205:HC205,">0")*B4/2	=F*HF205/(F+(1-F)*EXP(HH205/ff))*100000	=(1-HD205)*((HF205-HF204)/4+HF204/2)+HH204
206	100.5	=SUMIF(F206:HC206,">0")*B4/2	=F*HF206/(F+(1-F)*EXP(HH206/ff))*100000	=(1-HD206)*((HF206-HF205)/4+HF205/2)+HH205
207	101	=SUMIF(F207:HC207,">0")*B4/2	=F*HF207/(F+(1-F)*EXP(HH207/ff))*100000	=(1-HD207)*((HF207-HF206)/4+HF206/2)+HH206
208	101.5	=SUMIF(F208:HC208,">0")*B4/2	=F*HF208/(F+(1-F)*EXP(HH208/ff))*100000	=(1-HD208)*((HF208-HF207)/4+HF207/2)+HH207
209	102	=SUMIF(F209:HC209,">0")*B4/2	=F*HF209/(F+(1-F)*EXP(HH209/ff))*100000	=(1-HD209)*((HF209-HF208)/4+HF208/2)+HH208
210	102.5	=SUMIF(F210:HC210,">0")*B4/2	=F*HF210/(F+(1-F)*EXP(HH210/ff))*100000	=(1-HD210)*((HF210-HF209)/4+HF209/2)+HH209

Fit

Fig. 51: MAXIMUM LIKELIHOOD TEMPLATE: Fitting worksheet (Calculation of differences - errors, between observed and calculated OBS(h,t)) w/ formulae

Column HI refers back to the Raw Data worksheet to acquire the observed data. We have used the ISBLANK construct to detect age groups for which no data had been typed in. The reason we do so is that in Excel™, a formula that refers to a blank cell, assumes a value of 0. What we want is for this point to be ignored, so we turn the mortality rate to a space, " ". Any formula that now uses this cell will give an error message instead. (Notice Cell HJ210 in Figure 49, contains the error message of #VALUE! since no observed data had been typed in for age 102.5, Cell B22 in Figure 42). When the sum of the errors is calculated in C23 (Figure 44), this error message is ignored because we used the SUMIF instead of the SUM command.

Column HJ then calculates the error (given a weighing factor, See Section 3.6.3), comparing the calculated OBS(h,t) to the raw data.



Microsoft Excel - N2M1 Formulas

File Edit View Insert Format Tools Data Window Help

	E	H	HJ	HK	HL
1					
2					
3					
4	Age at Initiation (a)→ Age at death (t) ↓	Data	Error		
181	88				
182	88.5				
183	89				
184	89.5				
185	90				
186	90.5				
187	91				
188	91.5				
189	92				
190	92.5	=IF(ISBLANK('Raw Data'!B20)," ",'Raw Data'!B20)	=(HG190-HI190)^2/HG190*'Raw Data'!C20		
191	93				
192	93.5				
193	94				
194	94.5				
195	95				
196	95.5				
197	96				
198	96.5				
199	97				
200	97.5	=IF(ISBLANK('Raw Data'!B21)," ",'Raw Data'!B21)	=(HG200-HI200)^2/HG200*'Raw Data'!C21		
201	98				
202	98.5				
203	99				
204	99.5				
205	100				
206	100.5				
207	101				
208	101.5				
209	102				
210	102.5	=IF(ISBLANK('Raw Data'!B22)," ",'Raw Data'!B22)	=(HG210-HI210)^2/HG210*'Raw Data'!C22		

Fit

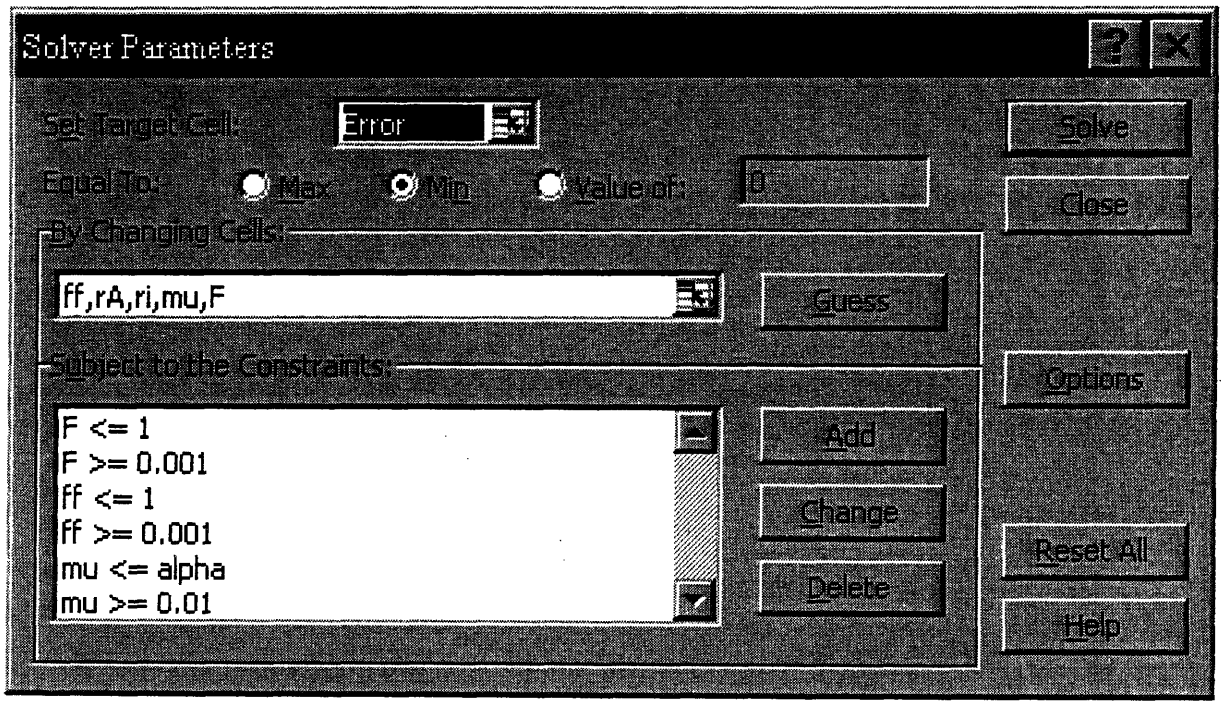
3.6.3.1 Determining Parameters for Best Fit by Maximum Likelihood

The template calculates mortality rates for a given set of parameters in fewer than 1 second, making this a potentially strong tool for maximum likelihood fits which use repeated iterations. The built-in maximum likelihood function is called the Solver.

Instructions:

1) Click on Fitting worksheet of template file (Figure 43)

2) Go to the Tools Menu and click on Solver (this feature is sometimes not included when Excel™ was installed; to install, insert Excel™ CD and select Custom Installation. This will give the option to add features that were not originally installed. The Solver is found under the Add-Ons section). The following window will come up.



3) To fit the data by maximum likelihood, the sum of the errors needs to be minimized, so select Cell C24 (given the name Error - Figure 43) as our Target cell, and click on Min (for minimize value). Set Changing Cells to the parameters of interest (to keep a parameter constant, simply do not include it in this list).

4) The only necessary constraints are the ones shown above, $(0.001 < F < 1)$, $(0.001 < f < 1)$, and $0.01 > (\alpha - \beta) > \alpha$. Additionally, under the Options section, there is a checkbox that constraints parameters to non-negative values.

5) Simply click on Solve to run the program.

The Solver will in time give a solution. Cells for the corresponding parameters will have been changed within the template by the Solver.

3.6.3.2 Effect of a Change in Parameter on Cancer Mortality Rates

The template can also be useful for demonstrative purposes, by illustrating each parameter's contribution to the carcinogenesis model, and how it affects the calculated mortality rates. Figure 52 summarizes the results one parameter at a time: the parameter (r_i r_j), divided by 2, 4, or 8; the parameter r_A divided by 2, 4, or 8; the parameter $(\alpha - \beta)$ multiplied by 0.9, 0.8, or 0.7; the parameter F multiplied by 0.9, 0.8, or 0.7; and the parameter 'f' multiplied by 1.2, 0.8, or 0.4.

By decreasing the initiation mutation rates, the magnitude of the 'slope' of the mortality curve decreases. The downwards trend in the slope of the curves is due to the slower accumulation of cells containing the first initiation mutation ($2 \tau r_i N_{max} a$). If the number of

these mutants decreases, there then is a smaller likelihood of initiation occurring at each age such that the overall slope of the curve rises slower.

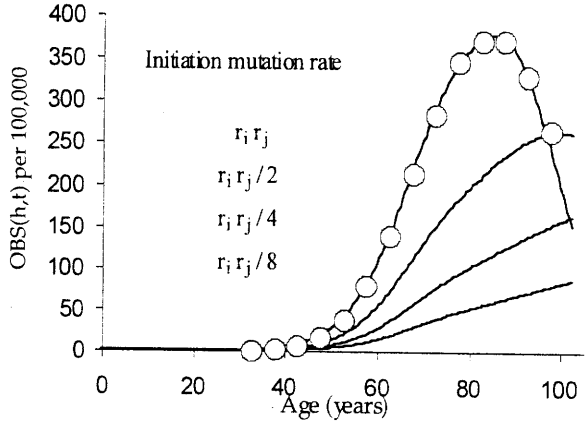
By decreasing the promotion mutation rates, the mortality curves shift rightwards. The lower the rate of promotion, the smaller the expected number of cancerous cells at age t. Since the probability of promotion is dependent on the expected number of cancerous cells, lowering the promotion mutation rate has the net effect of delaying death.

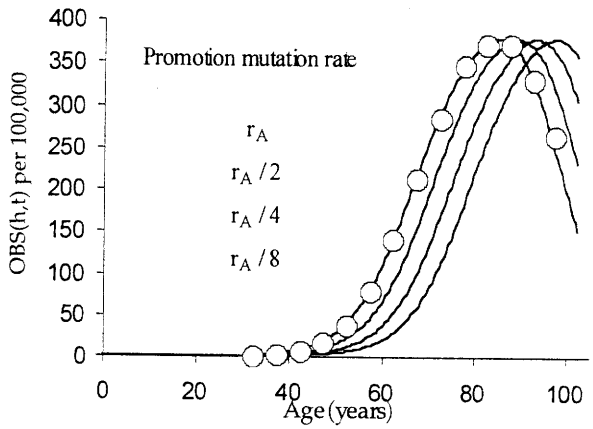
Decreasing the growth rate of precancerous lesions has a similar effect to a decrease in both promotion and initiation mutation rates, as the expected mortality curves are shifted both rightwards and downwards. As was the case when lowering the rate of promotion, decreasing the precancerous growth rate decreases the expected number of cancerous cells at age t (given that there are fewer precancerous cells at age t). Since the probability of promotion is dependent on the number of cancerous cells, lowering the growth rate has the net effect of delaying death (shift rightwards). Additionally, a slower growing adenoma is more likely to become extinct, where the probability of survival had been previously defined as $\left(\frac{\alpha-\beta}{\alpha}\right)$, (Moolgavkar et al, 1990b). Since fewer lesions are surviving stochastic extinction, then the mortality rates are consequently decreased as the production of new precancerous lesions with age is decreased.

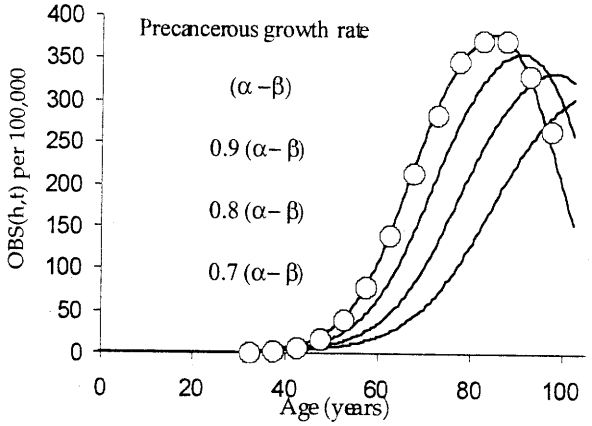
Decreasing the fraction at risk has the net effect of decreasing both the slope and area of the mortality curves. The area is decreased, as quite logically, there are fewer individuals who can die of the disease. By the same reasoning, since at each age there are fewer individuals at risk, then there are fewer people at every age that actually die (observed mortality rate is #dead/#alive) thereby decreasing the slope.

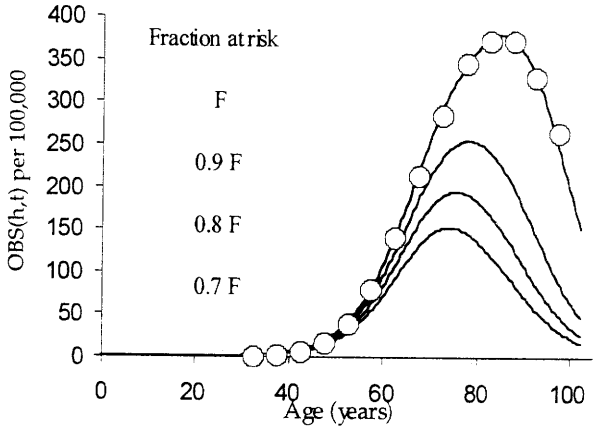
Fig. 52 Effect of change in a parameter on cancer mortality rates.

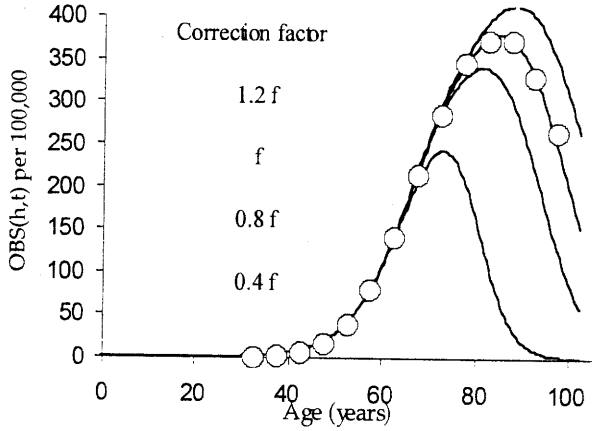
Effect of change in one of the parameters ($(r_i\ r_j)$, r_A, $(\alpha - \beta)$, F, and f) on the cancer mortality rates.


Initiation mutation rate
$r_i r_j$
$r_i r_j / 2$
$r_i r_j / 4$
$r_i r_j / 8$
OBS(h,t) per 100,000
Age (years)


Promotion mutation rate
r_A
$r_A / 2$
$r_A / 4$
$r_A / 8$
OBS(h,t) per 100,000
Age (years)


Precancerous growth rate
$(\alpha - \beta)$
$0.9\ (\alpha - \beta)$
$0.8\ (\alpha - \beta)$
$0.7\ (\alpha - \beta)$
OBS(h,t) per 100,000
Age (years)


Fraction at risk
F
0.9 F
0.8 F
0.7 F
OBS(h,t) per 100,000
Age (years)


Correction factor
1.2 f
f
0.8 f
0.4 f
OBS(h,t) per 100,000
Age (years)

To explain the effect of changing $\mathbf{f_h}$ on cancer mortality curves requires a little more explanation. This factor determines in part whether an individual at risk for the cancer of interest is still alive. As a result, the factor $\mathbf{f_h}$ primary plays a role in the elderly population, as these individuals are more likely to have already died of a connected form of death. It has a similar effect to changing the fraction at risk, $\mathbf{F_h}$, in that it determines how many people actually die of the cancer of interest; in other words, an individual cannot die of the cancer of interest if they already have died of a connected form of death. Decreasing $\mathbf{f_h}$ therefore increases the relative risk of dying from a connected form of death instead, thereby decreasing the area under the mortality curves for the cancer of interest. It however does not have an effect on the slope of the mortality curves, because the majority of the individuals who are at risk for dying of the cancer have not yet died of the cancer or any connected forms of death during these ages. Changing $\mathbf{f_h}$ has the net effect of determining when the mortality curve peaks.

3.6.4 Five Equations for Five Unknowns $\mathbf{F_h}$, $\mathbf{f_h}$, $\mathbf{r_i}$, $\mathbf{r_A}$, and $(\alpha - \beta)$ Template

Because of the complexity of the three-stage carcinogenesis model, maximum likelihood techniques can give different final solutions (depending on initial parameter values). To avoid this complication, Sections 3.5.4 and 3.5.5 formulated a methodology by which to explicitly calculate the parameters for the three-stage carcinogenesis model without the need to use maximum likelihood techniques. This section describes the construction of a template that takes full advantage of this methodology to estimate the parameters of the carcinogenesis model. (Figures 53-59)

Instructions on how to use this template are included in the legend to the figures.

Fig. 53: 5 EQUATIONS 5 UNKNOWNS TEMPLATE: Raw Data worksheet

Instructions:

1) Transcribe mortality data adjusted for underreporting, $OBS^{R}(h,t)$, into Column B, by appropriate age category.

2) Transcribe survival data, S(h,t), into Column C, by appropriate age category.

3) The plots reveal the estimates for the values of Δ_h and $(\alpha - \beta)$

(plots are of the Determine Δ column versus Age column,

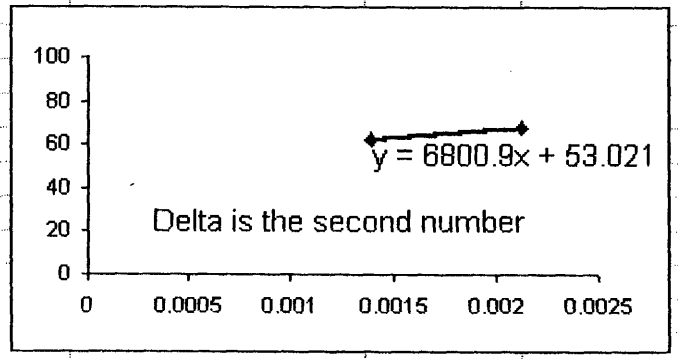
and the Determine $(\alpha - \beta)$ column versus Age column respectively)

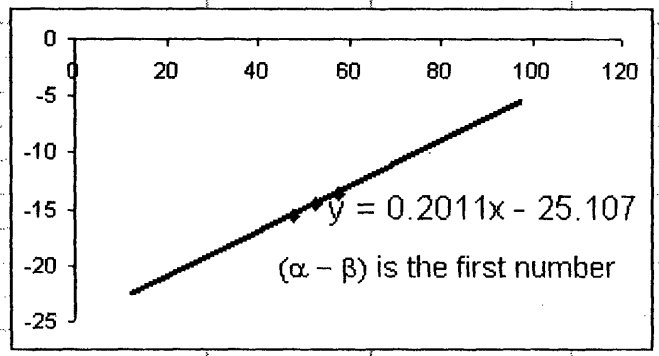
Formulae are shown in Figure 54

Microsoft Excel - Fit Program

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F	G	H	I
1	Age	OBS(t)/R(t) per 100,000	S(t)	Data	Slope	Determine (α - β)	Determine Δ	Determine area	
2	3		0.00	0	0				
3	7.5		0.00	0	0				
4	12.5		0.00	0	0			0	
5	17.5		0.00	0	0			0	
6	22.5		0.00	0	0			0	
7	27.5		0.00	0	0			0	
8	32.5	0.892303588	0.00	0.892303588	1.78461E-06			6.692276907	
9	37.5	2.318752323	0.00	2.318752323	2.8529E-06			8.027639778	
10	42.5	6.102864124	0.00	6.102864124	7.56822E-06			21.05404112	
11	47.5	16.22427966	0.00	16.22427966	2.02428E-05	-15.5922294		55.81785945	Area
12	52.5	38.07965624	0.00	38.07965624	4.37108E-05	-14.48165224		135.7598397	0.16423095
13	57.5	78.86345626	0.00	78.86345626	8.15676E-05	-13.58164427		292.3577813	
14	62.5	139.3796775	0.00	139.3796775	0.000121032		0.0013938	545.6078343	
15	67.5	212.8998573	0.00	212.8998573	0.00014704		0.002129	880.698837	
16	72.5	283.7833627	0.00	283.7833627	0.000141767			1241.70805	
17	77.5	345.9750636	0.00	345.9750636	0.000124383			1574.396066	
18	82.5	370.7324432	0.00	370.7324432	4.95148E-05			1791.768767	
19	87.5	371.5141656	0.00	371.5141656	1.56344E-06				
20	92.5	329.3603306	0.00	329.3603306	-8.43077E-05				
21	97.5	264.3553957	0.00	264.3553957	-0.00013001				
22	102.5		0.00	0	-0.000528711				
23									
24					0.000144404				
25									
26									
27									
28									
29									
30									
31									
32									
33									





Raw Data | Fitting

Fig. 54: 5 EQUATIONS 5 UNKNOWNS TEMPLATE: Raw Data worksheet w/ formulae

Formulae:

Column D calculates OBS*(h,t) by dividing the values from Column B, OBS^R (h,t) by the values from Column C, (1 – S(h,t)). (See Figure 53)

Column E calculates the derivative of OBS*(h,t) approximating it as $\Delta OBS^*(h,t) / \Delta t$
(Mortality rates were expressed per 100,000 so column divides back 100,000 to calculate the derivative)
INSTRUCTION: Transcribe the value for the maximum slope in Cell E24. In the case that there are two similar slopes, take the average (as illustrated)

Column F calculates the $\log_2$ of the derivative of OBS*(h,t).
INSTRUCTION: Fill in all cells with the formula for all ages 17.5 to 57.5. This creates a line with slope ($\alpha - \beta$) as shown in plot (Figure 53). Remove the formula from any cell corresponding to a point of this plot that does not lie on the line.

Column G helps calculate Δ_h.
INSTRUCTION: Fill in the formula only for the cells corresponding to the two age groups that encompass the maximum slope. The x-intercept of a line drawn through these two points estimates Δ_h.

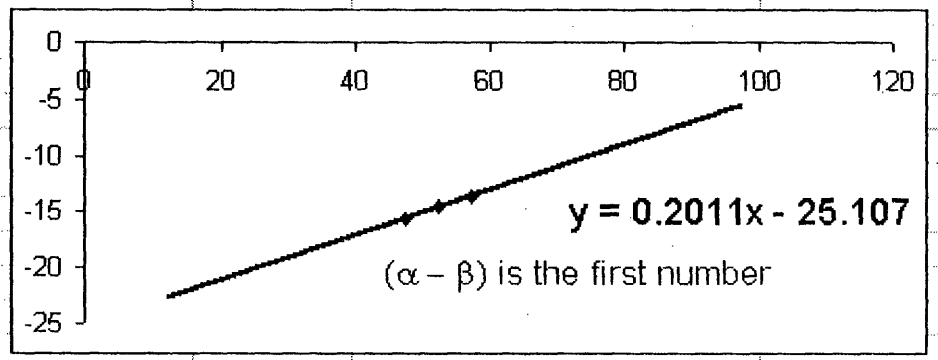
Column H helps calculate A_h by taking the integral of the mortality data OBS^R(h,t) in Column B. Integral done by trapezoidal approximation.
INSTRUCTION: Fill in the formula only for the cells <u>prior</u> to the maximum. Note that the maximum can occur between two points. In this mortality curve, the rates are similar for the age groups 82.5 and 87.5 (Figure 53), so the maximum must occur between these two points.

This however calculates the area only up to the last cell with a formula. It is tempting to assume that mortality curves are symmetrical, in which case we would multiply by 2, but to verify this we would need a mortality curve that goes up to age **160** or possibly more. Alternately, this template simply calculates expected mortality curves for ages up to 200, and then determines the expected symmetry of the curve and the percentage of the area before and after the maximum. Cell I11 multiplies the sum of the areas by this symmetry factor (Reference Fitting!L4, See Figure 58) to estimate the actual area. Again this is divided by 100,000 since the mortality rates had been expressed per 100,000.

Microsoft Excel - Fit Program

File Edit View Insert Format Tools Data Window Help

	D	E	F	G	H	I
1	Data	Slope	Determine (α – β)	Determine Δ	Determine area	
2	=B2/(1-C2)	=(D2)/3*10^(-5)				
3	=B3/(1-C3)	=(D3-D2)/4.5*10^(-5)				
4	=B4/(1-C4)	=(D4-D3)/5*10^(-5)			=(B4-B3)*2.5+MIN(B3:B4)*5	
5	=B5/(1-C5)	=(D5-D4)/5*10^(-5)			=(B5-B4)*2.5+MIN(B4:B5)*5	
6	=B6/(1-C6)	=(D6-D5)/5*10^(-5)			=(B6-B5)*2.5+MIN(B5:B6)*5	
7	=B7/(1-C7)	=(D7-D6)/5*10^(-5)			=(B7-B6)*2.5+MIN(B6:B7)*5	
8	=B8/(1-C8)	=(D8-D7)/5*10^(-5)			=(B8-B7)*2.5+MIN(B7:B8)*5	
9	=B9/(1-C9)	=(D9-D8)/5*10^(-5)			=(B9-B8)*2.5+MIN(B8:B9)*5	
10	=B10/(1-C10)	=(D10-D9)/5*10^(-5)			=(B10-B9)*2.5+MIN(B9:B10)*5	
11	=B11/(1-C11)	=(D11-D10)/5*10^(-5)	=LOG(E11,2)		=(B11-B10)*2.5+MIN(B10:B11)*5	Area
12	=B12/(1-C12)	=(D12-D11)/5*10^(-5)	=LOG(E12,2)		=(B12-B11)*2.5+MIN(B11:B12)*5	=Fitting!L4*SUM(H6:H21)*10^(-5)
13	=B13/(1-C13)	=(D13-D12)/5*10^(-5)	=LOG(E13,2)		=(B13-B12)*2.5+MIN(B12:B13)*5	
14	=B14/(1-C14)	=(D14-D13)/5*10^(-5)		=D14/10^5	=(B14-B13)*2.5+MIN(B13:B14)*5	
15	=B15/(1-C15)	=(D15-D14)/5*10^(-5)		=D15/10^5	=(B15-B14)*2.5+MIN(B14:B15)*5	
16	=B16/(1-C16)	=(D16-D15)/5*10^(-5)			=(B16-B15)*2.5+MIN(B15:B16)*5	
17	=B17/(1-C17)	=(D17-D16)/5*10^(-5)			=(B17-B16)*2.5+MIN(B16:B17)*5	
18	=B18/(1-C18)	=(D18-D17)/5*10^(-5)			=(B18-B17)*2.5+MIN(B17:B18)*5	
19	=B19/(1-C19)	=(D19-D18)/5*10^(-5)				
20	=B20/(1-C20)	=(D20-D19)/5*10^(-5)				
21	=B21/(1-C21)	=(D21-D20)/5*10^(-5)				
22	=B22/(1-C22)	=(D22-D21)/5*10^(-5)				
23						
24		=AVERAGE(E15:E16)				
25						
26						
27	x + 53.021					
28						
29						
30						
31						
32	0.0025					
33						



Raw Data | Fitting

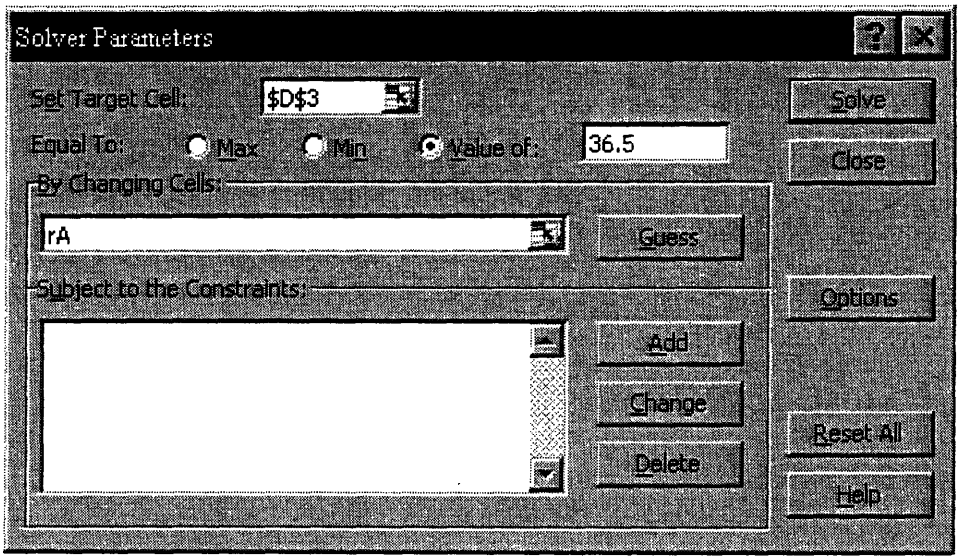
Fig. 55: 5 EQUATIONS 5 UNKNOWNS TEMPLATE: Fitting worksheet (Parameter section)

Columns A and C contain the name tags of the parameters used in the program.

Worksheet contains the calculated errors for the slope estimate, the area estimate, and the estimate of F.

This worksheet also contains a cell which converts the age to the row number of reference (i.e. age 5.5 corresponds to row 17).

Given a value of $(\alpha - \beta)$ and Δ_h, the Solver can be used to estimate the promotion mutation rate. Go to the Tools Menu and click on Solver (this feature is sometimes not included when Excel™ was installed; to install, insert Excel™ CD and select Custom Installation. This will give you the option to add features that were not originally installed. The Solver is found under the Add-Ons section). The following window will come up.

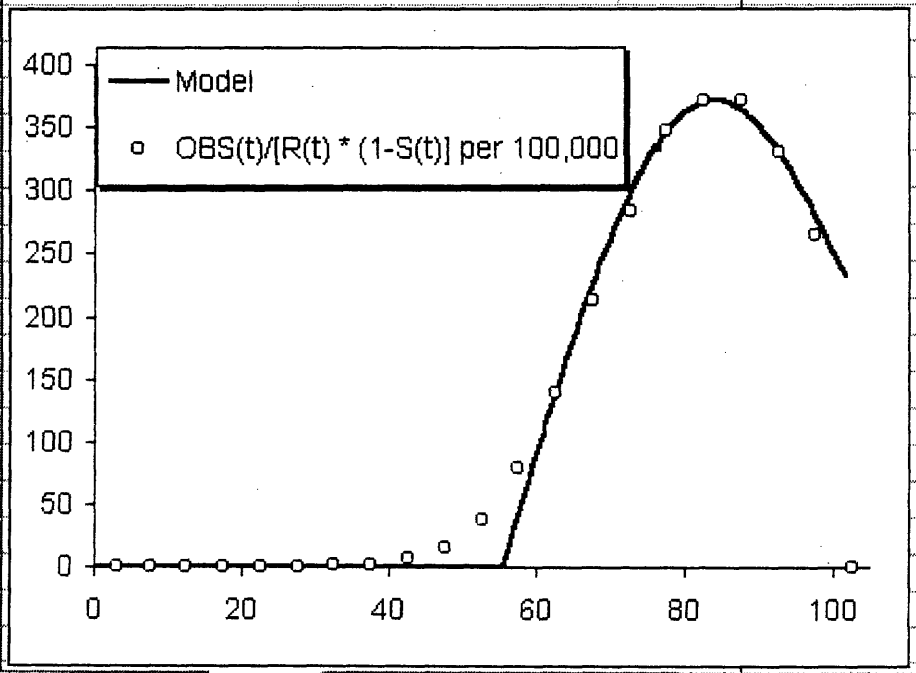


The Target Cell is the Calculated Δ (Cell D3). Select the value this cell should equal and select r_A as the parameter to change. Clicking on Solve will change r_A until the calculated Δ equals the value typed in.

Microsoft Excel - Fit Program

File Edit View Insert Format Tools Data Window Help

	A	B	C	D	E	F	G	H
1	N_{max}	2.40E+08			Age	Row	Row	Age
2	τ	5.7	$ri * 10^5 = rj / 3$	5.58E-02	5.5	17	30	12.5
3	$(\alpha_c - \beta_c)/\alpha_c$	0.3365	Calculated Δ	36.50	67.5	141	174	84.5
4	$2\tau^2 r_i r_j N_{max} (\alpha - \beta)/\alpha$	0.0004	Δ	55.50				
5	α	12.7	tmax	82.5			S(t)	t = age at death
6	$r_A * 10^5$	2.60E-02					0	0
7	$\alpha - \beta$	0.3232					0	0.5
8	F	0.5385					0	1
9	f	0.1957					0	1.5
10							0	2
11							0	2.5
12	Error		Sum Errors				0	3
13	Area	8.52E-02	8.52E-02				0	3.5
14	F	6.49E-06					0	4
15	Slope	4.39E-07					0	4.5
16							0	5
17							0	5.5
18							0	6
19							0	6.5
20							0	7
21							0	7.5
22							0	8
23							0	8.5
24							0	9
25							0	9.5
26							0	10
27							0	10.5
28							0	11
29							0	11.5
30							0	12
31							0	12.5
32							0	13
33							0	13.5
34							0	14



Raw Data | Fitting

Fig. 56: 5 EQUATIONS 5 UNKNOWNS: Fitting worksheet (Parameter section) w/ formulae

Instructions:

The following cells must be filled in,

Cell B1: Number of cells in adult, N_{max}
Cell B2: Stochastic survival probability of cancerous cell, $(\alpha_c - \beta_c)/\alpha_c$
Cell B3: Normal cell turnover rate, τ
Cell B5: Precancerous cell division rate, α
Cell B7: Growth rate of precancerous lesion, $\alpha - \beta$
Cell B8: Fraction at risk, F
Cell B9: Correction factor for connected forms of death, f

Formulae

Cell B4: Parameter equivalent to K_h

Error cells
Cell B13: Absolute value of (observed area – expected area)/expected area
Observed area – as calculated in Cell I12 of Raw Data worksheet (Figure 54)
Expected area – Equation 23, Section 3.5.4.2
Cell B14: Absolute value of (F – expected F)/F
Expected F – Equation 24, Section 3.5.4.2
Cell B15: Absolute value of (observed maximum slope – expected maximum slope) / expected maximum slope
Observed slope – as calculated in Cell E24 of Raw Data worksheet (Figure 54)
Expected maximum slope – F K_h
The maximum slope could be also described by evaluating the derivative of OBS*(h,t) at t = age at which the slope is maximum. Section 3.5.4.2 describes how to calculate this derivative. The resulting formula for this is shown in Cell B15.

Instructions: Equation 24 (Section 3.5.4.2) has an exponential term for the integral of the product of $P_{OBS}(h,t)$ $(1 - S(h,t))$ evaluated at t_{max}. The column for this calculation is K (Figure 58). The row number varies by age; the appropriate row can be calculated by using the feature as mentioned in Figure 55 to convert ages into row numbers. Put in proper row number in cell B14.

If using the alternate maximum slope method, the integral P_{OBS} (h,t) $(1 - S(h,t))$ evaluated at t = age of maximum slope is needed. Put in proper row number in cell B15.

Microsoft Excel - Fit Program

File Edit View Insert Format Tools Data Window Help

	A	B
1	N_{max}	240000000
2	τ	5.7
3	$(\alpha_c - \beta_c)/\alpha_c$	=(42.2-28)/42.2
4	$2\tau^2 r_i r_j N_{max} (\alpha - \beta)/\alpha$	=2*tau^2*(ri/100000)^2*3*Nmax*mu/alpha
5	α	12.7
6	$r_A * 10^5$	0.026
7	$\alpha - \beta$	0.3232
8	F	0.538467219868856
9	f	0.195725366517599
10		0
11		
12	Error	
13	Area	=ABS('Raw Data'!I12-(-ff*LN(1-F)))/(-ff*LN(1-F))
14	F	=ABS(((ff-2*K175)/(ff*(1-EXP(-K175/ff))-2*K175)-F)/(ff-2*K175)/(ff*(1-EXP(-K175/ff))-2*K175))
15	Slope	=ABS(((EXP(-K146/ff)*F*kappa*(ff*F*EXP(-K146/ff)+(1-F)*(ff*(1-K146*2/ff))))/(ff*(1-F*(1-EXP(-K146/ff)))^2))-'Raw Data'!E24)/((EXP(-
16		K146/ff)*F*kappa*(ff*F*EXP(-K146/ff)+(1-F)*(ff*(1-K146*2/ff))))/(ff*(1-F*(1-EXP(-K146/ff)))^2))

Raw Data | Fitting

Fig. 57: 5 EQUATIONS 5 UNKNOWNS: Fitting worksheet - $P_{OBS}(h,t)$ w/ formulae

Worksheet calculates $P_{OBS}(h,t)$ (Column I) and $P_{OBS}(h,t)$ $(1 - S(h,t))$ (Column J) based on the modified Nordling model (Equation 19, Section 3.5.4)

$$\mathbf{P_{OBS}(h,t)} = \mathbf{K_h}\,(t - \Delta_h) \quad (t > \Delta_h)$$

Microsoft Excel - Fit Program

File Edit View Insert Format Tools Data Window Help

	G	H	I	J
1	Row	Age		
2	30	=(G2-5)/2		
3	174	=(G3-5)/2		
4				
5	S(t)	t = age at death	P(t)	(1-S(t))*P(t)
6	0	0	=IF(H6>delta,kappa*(H6-delta),0)	=I6*(1-G6)
7	0	0.5	=IF(H7>delta,kappa*(H7-delta),0)	=I7*(1-G7)
8	0	1	=IF(H8>delta,kappa*(H8-delta),0)	=I8*(1-G8)
9	0	1.5	=IF(H9>delta,kappa*(H9-delta),0)	=I9*(1-G9)
10	0	2	=IF(H10>delta,kappa*(H10-delta),0)	=I10*(1-G10)
11	0	2.5	=IF(H11>delta,kappa*(H11-delta),0)	=I11*(1-G11)
12	0	3	=IF(H12>delta,kappa*(H12-delta),0)	=I12*(1-G12)
13	0	3.5	=IF(H13>delta,kappa*(H13-delta),0)	=I13*(1-G13)
14	0	4	=IF(H14>delta,kappa*(H14-delta),0)	=I14*(1-G14)
15	0	4.5	=IF(H15>delta,kappa*(H15-delta),0)	=I15*(1-G15)
16	0	5	=IF(H16>delta,kappa*(H16-delta),0)	=I16*(1-G16)
17	0	5.5	=IF(H17>delta,kappa*(H17-delta),0)	=I17*(1-G17)
18	0	6	=IF(H18>delta,kappa*(H18-delta),0)	=I18*(1-G18)
19	0	6.5	=IF(H19>delta,kappa*(H19-delta),0)	=I19*(1-G19)
20	0	7	=IF(H20>delta,kappa*(H20-delta),0)	=I20*(1-G20)
21	0	7.5	=IF(H21>delta,kappa*(H21-delta),0)	=I21*(1-G21)
22	0	8	=IF(H22>delta,kappa*(H22-delta),0)	=I22*(1-G22)
23	0	8.5	=IF(H23>delta,kappa*(H23-delta),0)	=I23*(1-G23)
24	0	9	=IF(H24>delta,kappa*(H24-delta),0)	=I24*(1-G24)
25	0	9.5	=IF(H25>delta,kappa*(H25-delta),0)	=I25*(1-G25)
26	0	10	=IF(H26>delta,kappa*(H26-delta),0)	=I26*(1-G26)
27	0	10.5	=IF(H27>delta,kappa*(H27-delta),0)	=I27*(1-G27)
28	0	11	=IF(H28>delta,kappa*(H28-delta),0)	=I28*(1-G28)
29	0	11.5	=IF(H29>delta,kappa*(H29-delta),0)	=I29*(1-G29)
30	0	12	=IF(H30>delta,kappa*(H30-delta),0)	=I30*(1-G30)
31	0	12.5	=IF(H31>delta,kappa*(H31-delta),0)	=I31*(1-G31)
32	0	13	=IF(H32>delta,kappa*(H32-delta),0)	=I32*(1-G32)
33	0	13.5	=IF(H33>delta,kappa*(H33-delta),0)	=I33*(1-G33)
34	0	14	=IF(H34>delta,kappa*(H34-delta),0)	=I34*(1-G34)
35	0	14.5	=IF(H35>delta,kappa*(H35-delta),0)	=I35*(1-G35)

Raw Data | Fitting

Fig. 58: 5 EQUATIONS 5 UNKNOWNS: Fitting worksheet - OBS(h,t) w/ formulae

Worksheet calculates the integral of $P_{OBS}(h,t)$ $(1 - S(h,t))$ using integration by trapezoids, and OBS(h,t) (Equation 16, Section 3.5.2.2)

$$OBS(h,t) = \frac{\mathbf{F_h} \cdot (1 - S(h,t)) \cdot R(h,t) \cdot \mathbf{P_{OBS}(h,t)}}{\mathbf{F_h} + (1 - \mathbf{F_h}) \cdot e^{\frac{1}{f_h}\int_0^t \mathbf{P_{OBS}(h,t)}\,(1 - S(h,t))\,dt}}$$

There are two other cells with formulas here. They are used to help calculate the symmetry of the mortality curve. It requires the age of the last observed point before the maximum of the curve, and the other is the relative area under the curve after this point (Calculated as shown in Figure 59). This factor was used to estimate the area under the curve in Cell I11 (Figure 54).

Microsoft Excel - Fit Program

File Edit View Insert Format Tools Data Window Help

	K	L
1		Last point before max
2		=tmax
3		1/Fraction after max
4		=N424/(N424-M424)
5	INT((1-S(t))*P(t))	Model
6		
7	=MIN(J7,J6)*0.5+ABS(J7-J6)*0.25+K6	=(F*I7)/((1-G7)*(F+(1-F)*EXP(K7/ff)))*100000
8	=MIN(J8,J7)*0.5+ABS(J8-J7)*0.25+K7	=(F*I8)/((1-G8)*(F+(1-F)*EXP(K8/ff)))*100000
9	=MIN(J9,J8)*0.5+ABS(J9-J8)*0.25+K8	=(F*I9)/((1-G9)*(F+(1-F)*EXP(K9/ff)))*100000
10	=MIN(J10,J9)*0.5+ABS(J10-J9)*0.25+K9	=(F*I10)/((1-G10)*(F+(1-F)*EXP(K10/ff)))*100000
11	=MIN(J11,J10)*0.5+ABS(J11-J10)*0.25+K10	=(F*I11)/((1-G11)*(F+(1-F)*EXP(K11/ff)))*100000
12	=MIN(J12,J11)*0.5+ABS(J12-J11)*0.25+K11	=(F*I12)/((1-G12)*(F+(1-F)*EXP(K12/ff)))*100000
13	=MIN(J13,J12)*0.5+ABS(J13-J12)*0.25+K12	=(F*I13)/((1-G13)*(F+(1-F)*EXP(K13/ff)))*100000
14	=MIN(J14,J13)*0.5+ABS(J14-J13)*0.25+K13	=(F*I14)/((1-G14)*(F+(1-F)*EXP(K14/ff)))*100000
15	=MIN(J15,J14)*0.5+ABS(J15-J14)*0.25+K14	=(F*I15)/((1-G15)*(F+(1-F)*EXP(K15/ff)))*100000
16	=MIN(J16,J15)*0.5+ABS(J16-J15)*0.25+K15	=(F*I16)/((1-G16)*(F+(1-F)*EXP(K16/ff)))*100000
17	=MIN(J17,J16)*0.5+ABS(J17-J16)*0.25+K16	=(F*I17)/((1-G17)*(F+(1-F)*EXP(K17/ff)))*100000
18	=MIN(J18,J17)*0.5+ABS(J18-J17)*0.25+K17	=(F*I18)/((1-G18)*(F+(1-F)*EXP(K18/ff)))*100000
19	=MIN(J19,J18)*0.5+ABS(J19-J18)*0.25+K18	=(F*I19)/((1-G19)*(F+(1-F)*EXP(K19/ff)))*100000
20	=MIN(J20,J19)*0.5+ABS(J20-J19)*0.25+K19	=(F*I20)/((1-G20)*(F+(1-F)*EXP(K20/ff)))*100000
21	=MIN(J21,J20)*0.5+ABS(J21-J20)*0.25+K20	=(F*I21)/((1-G21)*(F+(1-F)*EXP(K21/ff)))*100000
22	=MIN(J22,J21)*0.5+ABS(J22-J21)*0.25+K21	=(F*I22)/((1-G22)*(F+(1-F)*EXP(K22/ff)))*100000
23	=MIN(J23,J22)*0.5+ABS(J23-J22)*0.25+K22	=(F*I23)/((1-G23)*(F+(1-F)*EXP(K23/ff)))*100000
24	=MIN(J24,J23)*0.5+ABS(J24-J23)*0.25+K23	=(F*I24)/((1-G24)*(F+(1-F)*EXP(K24/ff)))*100000
25	=MIN(J25,J24)*0.5+ABS(J25-J24)*0.25+K24	=(F*I25)/((1-G25)*(F+(1-F)*EXP(K25/ff)))*100000
26	=MIN(J26,J25)*0.5+ABS(J26-J25)*0.25+K25	=(F*I26)/((1-G26)*(F+(1-F)*EXP(K26/ff)))*100000
27	=MIN(J27,J26)*0.5+ABS(J27-J26)*0.25+K26	=(F*I27)/((1-G27)*(F+(1-F)*EXP(K27/ff)))*100000
28	=MIN(J28,J27)*0.5+ABS(J28-J27)*0.25+K27	=(F*I28)/((1-G28)*(F+(1-F)*EXP(K28/ff)))*100000
29	=MIN(J29,J28)*0.5+ABS(J29-J28)*0.25+K28	=(F*I29)/((1-G29)*(F+(1-F)*EXP(K29/ff)))*100000
30	=MIN(J30,J29)*0.5+ABS(J30-J29)*0.25+K29	=(F*I30)/((1-G30)*(F+(1-F)*EXP(K30/ff)))*100000
31	=MIN(J31,J30)*0.5+ABS(J31-J30)*0.25+K30	=(F*I31)/((1-G31)*(F+(1-F)*EXP(K31/ff)))*100000
32	=MIN(J32,J31)*0.5+ABS(J32-J31)*0.25+K31	=(F*I32)/((1-G32)*(F+(1-F)*EXP(K32/ff)))*100000
33	=MIN(J33,J32)*0.5+ABS(J33-J32)*0.25+K32	=(F*I33)/((1-G33)*(F+(1-F)*EXP(K33/ff)))*100000
34	=MIN(J34,J33)*0.5+ABS(J34-J33)*0.25+K33	=(F*I34)/((1-G34)*(F+(1-F)*EXP(K34/ff)))*100000
35	=MIN(J35,J34)*0.5+ABS(J35-J34)*0.25+K34	=(F*I35)/((1-G35)*(F+(1-F)*EXP(K35/ff)))*100000

Raw Data | Fitting

Fig. 59: 5 EQUATIONS 5 UNKNOWNS: Fitting worksheet - Area w/ formulae

Worksheet determines symmetry of the mortality curve by calculating area (integration by trapezoids) of the area before the maximum, and finding what fraction of the total area it comprises.

Microsoft Excel - Fit Program

File Edit View Insert Format Tools Data Window Help

	M	N
1		
2		
3		
4		
5	Area after max	Total area
6		
7	=IF(H8>=L2,MIN(L8,L7)*0.5+ABS(L8-L7)*0.25)+M6	=MIN(L8,L7)*0.5+ABS(L8-L7)*0.25+N6
8	=IF(H9>=L2,MIN(L9,L8)*0.5+ABS(L9-L8)*0.25)+M7	=MIN(L9,L8)*0.5+ABS(L9-L8)*0.25+N7
9	=IF(H10>=L2,MIN(L10,L9)*0.5+ABS(L10-L9)*0.25)+M8	=MIN(L10,L9)*0.5+ABS(L10-L9)*0.25+N8
10	=IF(H11>=L2,MIN(L11,L10)*0.5+ABS(L11-L10)*0.25)+M9	=MIN(L11,L10)*0.5+ABS(L11-L10)*0.25+N9
11	=IF(H12>=L2,MIN(L12,L11)*0.5+ABS(L12-L11)*0.25)+M10	=MIN(L12,L11)*0.5+ABS(L12-L11)*0.25+N10
12	=IF(H13>=L2,MIN(L13,L12)*0.5+ABS(L13-L12)*0.25)+M11	=MIN(L13,L12)*0.5+ABS(L13-L12)*0.25+N11
13	=IF(H14>=L2,MIN(L14,L13)*0.5+ABS(L14-L13)*0.25)+M12	=MIN(L14,L13)*0.5+ABS(L14-L13)*0.25+N12
14	=IF(H15>=L2,MIN(L15,L14)*0.5+ABS(L15-L14)*0.25)+M13	=MIN(L15,L14)*0.5+ABS(L15-L14)*0.25+N13
15	=IF(H16>=L2,MIN(L16,L15)*0.5+ABS(L16-L15)*0.25)+M14	=MIN(L16,L15)*0.5+ABS(L16-L15)*0.25+N14
16	=IF(H17>=L2,MIN(L17,L16)*0.5+ABS(L17-L16)*0.25)+M15	=MIN(L17,L16)*0.5+ABS(L17-L16)*0.25+N15
17	=IF(H18>=L2,MIN(L18,L17)*0.5+ABS(L18-L17)*0.25)+M16	=MIN(L18,L17)*0.5+ABS(L18-L17)*0.25+N16
18	=IF(H19>=L2,MIN(L19,L18)*0.5+ABS(L19-L18)*0.25)+M17	=MIN(L19,L18)*0.5+ABS(L19-L18)*0.25+N17
19	=IF(H20>=L2,MIN(L20,L19)*0.5+ABS(L20-L19)*0.25)+M18	=MIN(L20,L19)*0.5+ABS(L20-L19)*0.25+N18
20	=IF(H21>=L2,MIN(L21,L20)*0.5+ABS(L21-L20)*0.25)+M19	=MIN(L21,L20)*0.5+ABS(L21-L20)*0.25+N19
21	=IF(H22>=L2,MIN(L22,L21)*0.5+ABS(L22-L21)*0.25)+M20	=MIN(L22,L21)*0.5+ABS(L22-L21)*0.25+N20
22	=IF(H23>=L2,MIN(L23,L22)*0.5+ABS(L23-L22)*0.25)+M21	=MIN(L23,L22)*0.5+ABS(L23-L22)*0.25+N21
23	=IF(H24>=L2,MIN(L24,L23)*0.5+ABS(L24-L23)*0.25)+M22	=MIN(L24,L23)*0.5+ABS(L24-L23)*0.25+N22
24	=IF(H25>=L2,MIN(L25,L24)*0.5+ABS(L25-L24)*0.25)+M23	=MIN(L25,L24)*0.5+ABS(L25-L24)*0.25+N23
25	=IF(H26>=L2,MIN(L26,L25)*0.5+ABS(L26-L25)*0.25)+M24	=MIN(L26,L25)*0.5+ABS(L26-L25)*0.25+N24
26	=IF(H27>=L2,MIN(L27,L26)*0.5+ABS(L27-L26)*0.25)+M25	=MIN(L27,L26)*0.5+ABS(L27-L26)*0.25+N25
27	=IF(H28>=L2,MIN(L28,L27)*0.5+ABS(L28-L27)*0.25)+M26	=MIN(L28,L27)*0.5+ABS(L28-L27)*0.25+N26
28	=IF(H29>=L2,MIN(L29,L28)*0.5+ABS(L29-L28)*0.25)+M27	=MIN(L29,L28)*0.5+ABS(L29-L28)*0.25+N27
29	=IF(H30>=L2,MIN(L30,L29)*0.5+ABS(L30-L29)*0.25)+M28	=MIN(L30,L29)*0.5+ABS(L30-L29)*0.25+N28
30	=IF(H31>=L2,MIN(L31,L30)*0.5+ABS(L31-L30)*0.25)+M29	=MIN(L31,L30)*0.5+ABS(L31-L30)*0.25+N29
31	=IF(H32>=L2,MIN(L32,L31)*0.5+ABS(L32-L31)*0.25)+M30	=MIN(L32,L31)*0.5+ABS(L32-L31)*0.25+N30
32	=IF(H33>=L2,MIN(L33,L32)*0.5+ABS(L33-L32)*0.25)+M31	=MIN(L33,L32)*0.5+ABS(L33-L32)*0.25+N31
33	=IF(H34>=L2,MIN(L34,L33)*0.5+ABS(L34-L33)*0.25)+M32	=MIN(L34,L33)*0.5+ABS(L34-L33)*0.25+N32
34	=IF(H35>=L2,MIN(L35,L34)*0.5+ABS(L35-L34)*0.25)+M33	=MIN(L35,L34)*0.5+ABS(L35-L34)*0.25+N33
35	=IF(H36>=L2,MIN(L36,L35)*0.5+ABS(L36-L35)*0.25)+M34	=MIN(L36,L35)*0.5+ABS(L36-L35)*0.25+N34

Raw Data | Fitting

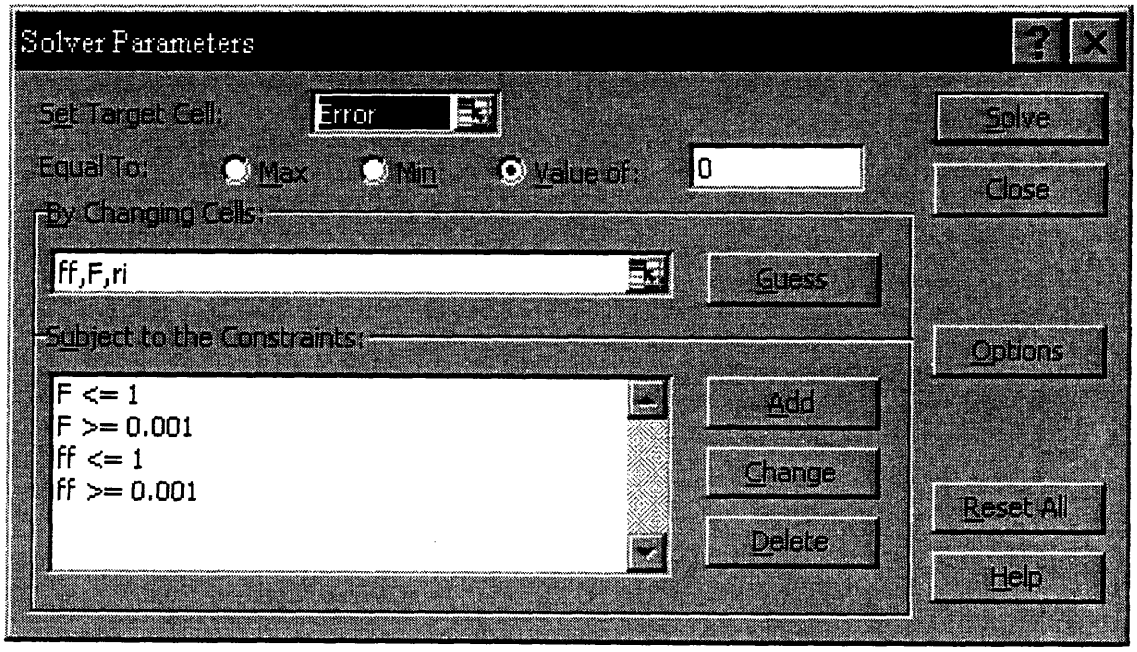
3.6.4.1 Determining Parameters by 5 Equations 5 Unknowns

The 5 Equations 5 Unknowns template described in Section 3.6.4 can further calculate $\mathbf{F_h}$, $\mathbf{f_h}$, and $\mathbf{r_i}$ by minimizing the error in the calculated area, calculated slope, and the calculated F given a t_{max}.

Instructions:

1) Click on Fitting worksheet of template file (Figure 55)

2) Go to the Tools Menu and click on Solver (this feature is sometimes not included when Excel™ was installed; to install, insert Excel™ CD and select Custom Installation. This will give you the option to add features that were not originally installed. The Solver is found under the Add-Ons section). The following window will come up.



6) To fit the three remaining features of the curve (slope, area, and maximum), the sum of the errors of these features needs to be minimized, so select Cell C13 (given the name Error - Figure 55) as our Target cell, and click on Min (for minimize value), or alternately click on Value of 0. Set Changing Cells to the parameters of interest (to keep a parameter constant, simply do not include it in this list).

7) The only necessary constraints are the ones shown above, ($0.001 < F < 1$) and ($0.001 < f < 1$). Additionally, under the Options section, there is a checkbox that constraints parameters to non-negative values.

8) Simply click on Solve to run the program.

The Solver will in time give a solution. Cells for the corresponding parameters will have been changed within the template by the Solver.

3.6.4.2 Caveats of the 5 Equations 5 Unknowns Methodology

Because mortality rates have been calculated in 5-year age groups, there is no guarantee that when selecting the age at which the mortality rate is maximum, and the age at which the slope of the mortality rates is maximum, that the right ages have been selected. Technically there is a 5-year window for these parameters based on observation alone.

Using the wrong ages along with the Solver in the 5 Equations 5 Unknowns template (Section 3.6.4.1) will give results that do not fit the observed mortality data. Adjusting them within their 5-year ranges and retrying the Solver will eventually give results that fit the data. Experience has though shown that the results do not vary by more than 5% even if the parameters chosen had been wrong.

A further caveat is that if we select the alternate method for calculating the expected maximum slope (See Figure 56 legend), then the observed estimate of Δ_h would be in error. As shown in Figure 60, because of the curvature of the observed mortality rate function, the x-intercept of the line going through the linear portion of OBS*(h,t) does not correspond with the expected value of Δ_h. Experience using this application has shown that the expected Δ_h is less than 5 years more than the observed Δ_h. When using the alternate method, we recommend first trying $\Delta_h + 2.5$ to get results that best fit the mortality data.

Fig. 60: Difference in the observed Δ_h and the expected Δ_h

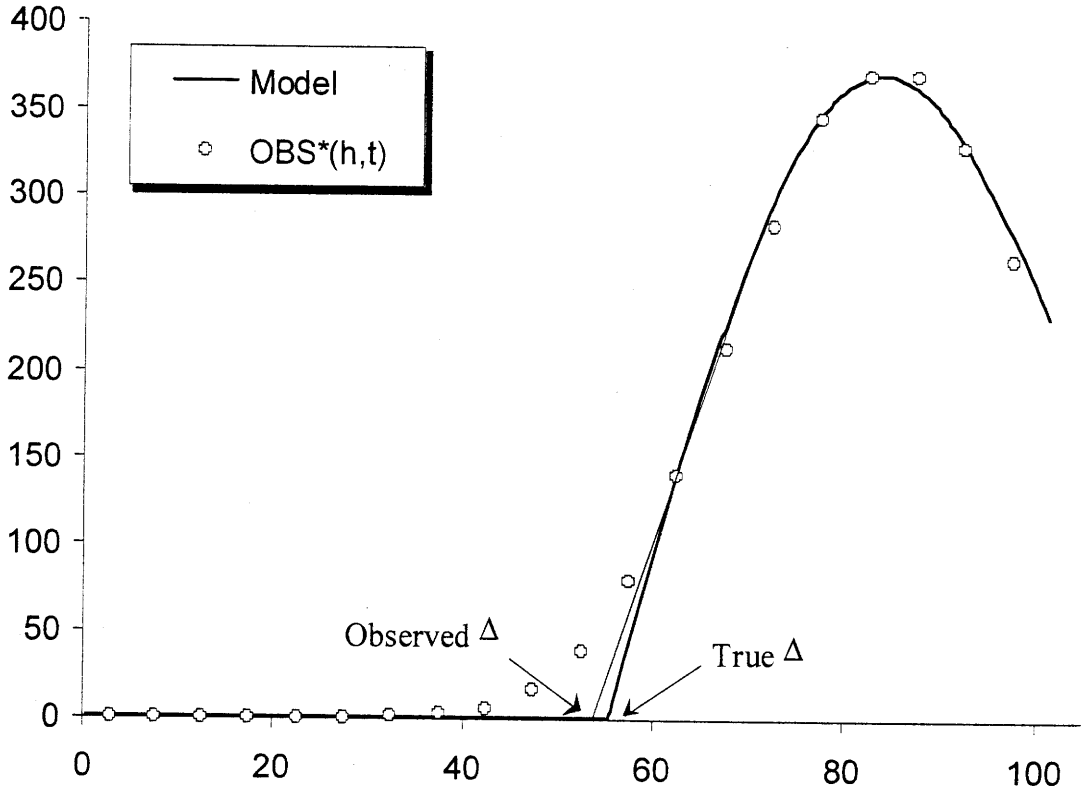
In the 5 equations 5 unknowns methodology, the approximation

$$OBS^*(h,t) = \frac{F_h \cdot \kappa_h \cdot (t - \Delta_h)}{F_h + (1 - F_h) \cdot e^{\frac{1}{f_h}\int_0^t \kappa_h \cdot (t - \Delta_h)(1 - S(h,t))\, dt}}$$

$$OBS^*(h,t) \sim F_h\, \kappa_h\, (t - \Delta_h) \qquad \text{for } t < t_{max}$$

was made.

However, the denominator is significant enough to create bending on the mortality curve even prior to t_{max}. As illustrated this affects the estimation of Δ_h by a few years.


Model
OBS*(h,t)
Observed Δ
True Δ
400
350
300
250
200
150
100
50
0
0
20
40
60
80
100

4. RESULTS

4.1 Colon Cancer

4.1.1 Calculated Parameters for the Case of (n = 2 and m = 1)

Table 2 and Figures 61, 62 summarize the results for the n=2, m=1 case, using the five equations five unknown methodology. Initiation mutation rate r_i values are reported assuming that $\mathbf{r_i} = \mathbf{r_j}/3$. We base this approximation on observations by Grist et al (1992) of HLA-A point mutations (r_i) and LOH (r_j) in presumptive T-cell stem cells *in vivo* and by de Nooij-van Dalen et al (1998) who studied the pathways of LOH at this same locus in human cells *in vitro*.

Table 2: Summary of primary and secondary risk parameters, (n = 2, m = 1).

	Birthyear	$\mathbf{F_h}$	$\mathbf{f_h}$	$\mathbf{r_i}$	$\mathbf{r_A}$	$\alpha - \beta$
EAM						
	1840s	0.30	0.11	4.4×10^{-8}	8.4×10^{-8}	--
	1850s	0.35	0.13	4.6×10^{-8}	8.8×10^{-8}	--
	1860s	0.38	0.15	5.6×10^{-8}	8.0×10^{-8}	--
	1870s	0.41	0.15	5.7×10^{-8}	8.2×10^{-8}	--
	1880s	0.40	0.21	5.9×10^{-8}	1.3×10^{-7}	0.19
	1890s	0.40	0.21	6.7×10^{-8}	8.6×10^{-8}	0.20
	1900s	0.39	0.24	7.0×10^{-8}	7.6×10^{-8}	0.21
	1910s	0.45	--	8.5×10^{-8}	8.1×10^{-8}	0.19
	1920s	0.43	--	7.6×10^{-8}	8.1×10^{-8}	0.21
	1930s	0.42	--	7.4×10^{-8}	8.1×10^{-8}	0.21
EAF						
	1840s	0.28	0.18	7.0×10^{-8}	1.6×10^{-7}	--
	1850s	0.33	0.18	7.0×10^{-8}	1.5×10^{-7}	--
	1860s	0.41	0.15	7.2×10^{-8}	1.2×10^{-7}	--
	1870s	0.39	0.15	7.3×10^{-8}	1.8×10^{-7}	--
	1880s	0.40	0.16	6.9×10^{-8}	2.5×10^{-7}	0.16
	1890s	0.40	0.17	7.0×10^{-8}	3.0×10^{-7}	0.16
	1900s	0.39	0.17	7.0×10^{-8}	2.6×10^{-7}	0.17
	1910s	0.39	--	7.2×10^{-8}	2.4×10^{-7}	0.17
	1920s	0.39	--	7.0×10^{-8}	2.6×10^{-7}	0.17
	1930s	0.39	--	7.0×10^{-8}	2.5×10^{-7}	0.17

Birthyear	F_h	f_h	r_i	r_A	$\alpha - \beta$
NEAM					
1850s	0.31	0.06	4.3×10^{-8}	5.8×10^{-8}	--
1860s	0.35	0.08	4.3×10^{-8}	5.8×10^{-8}	--
1870s	0.36	0.11	4.3×10^{-8}	7.1×10^{-8}	--
1880s	0.43	0.13	4.7×10^{-8}	8.5×10^{-8}	0.19
1890s	0.45	0.17	6.1×10^{-8}	7.4×10^{-8}	0.19
1900s	0.43	0.21	7.0×10^{-8}	8.1×10^{-8}	0.18
1910s	0.43	--	6.2×10^{-8}	6.8×10^{-8}	0.21
1920s	0.43	--	6.2×10^{-8}	5.6×10^{-8}	0.23
1930s	0.45	--	6.9×10^{-8}	8.2×10^{-8}	0.21
NEAF					
1860s	0.41	0.08	4.8×10^{-8}	8.0×10^{-8}	--
1870s	0.45	0.11	4.4×10^{-8}	1.5×10^{-7}	--
1880s	0.45	0.13	4.6×10^{-8}	2.0×10^{-7}	0.17
1890s	0.44	0.14	6.2×10^{-8}	2.0×10^{-7}	0.15
1900s	0.45	0.14	6.4×10^{-8}	2.5×10^{-7}	0.15
1910s	0.41	--	5.9×10^{-8}	2.6×10^{-7}	0.16
1920s	0.42	--	5.8×10^{-8}	2.3×10^{-7}	0.17
1930s	0.42	--	6.1×10^{-8}	1.8×10^{-7}	0.17

Comprehensive analysis of the data for birth year cohorts from 1910 onwards was not possible. Estimation of $\mathbf{f_h}$ was not possible for these birth year cohorts, as reasonable knowledge of how the mortality rates decrease at extreme old age is required. ($\alpha-\beta$) was still observable for more recent birthyear cohorts permitting calculation of the promotion mutation rate, $\mathbf{r_A}$. $\mathbf{F_h}$ and $\mathbf{r_i}$ were approximated by noting that the slope of the colon cancer rates for the later birth years did not show significant changes in their slope (Figures 31 and 32), suggesting that the area under the mortality curves would be constant. If future data demonstrates that the area actually changed, this will necessarily have been due to a result of a change in $\mathbf{f_h}$, but not in $\mathbf{F_h}$, as a change in $\mathbf{F_h}$ would have affected both the slope and area of the age-specific colon cancer mortality rate function (Figure 52).

Fig. 61: Historical trend in $\mathbf{F_h}$ and $\mathbf{r_i}$, for colon cancer

Historical trend in the calculated fraction at primary risk for colon cancer, $\mathbf{F_h}$, and the calculated initiation mutation rate, $\mathbf{r_i}$, for colon cancer.

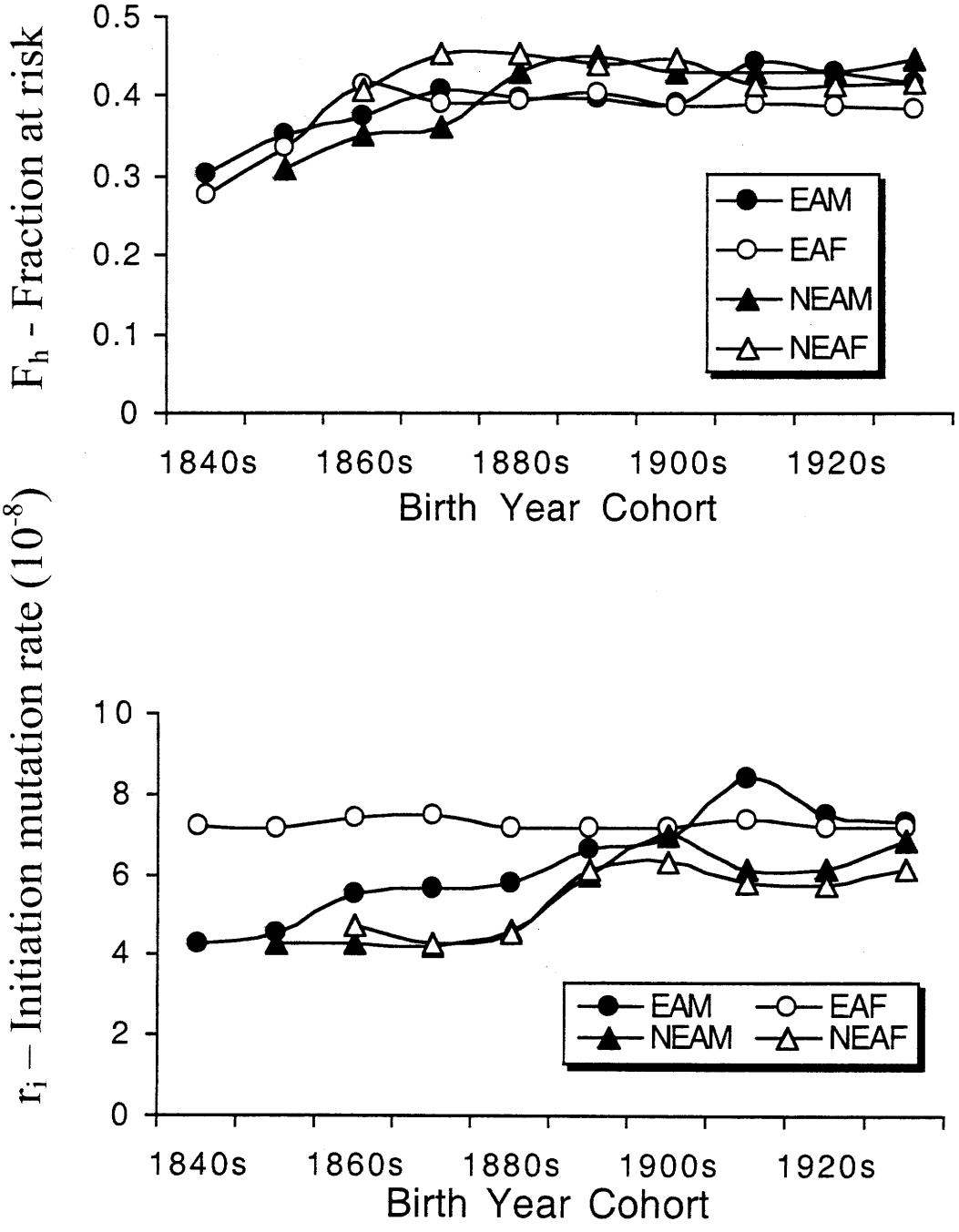
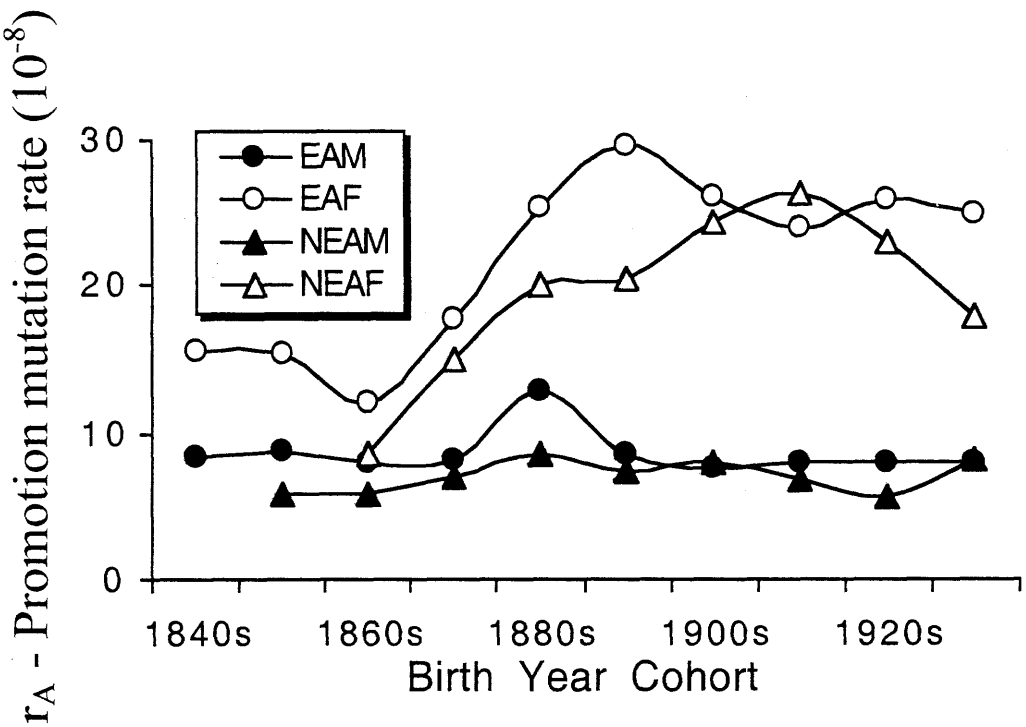
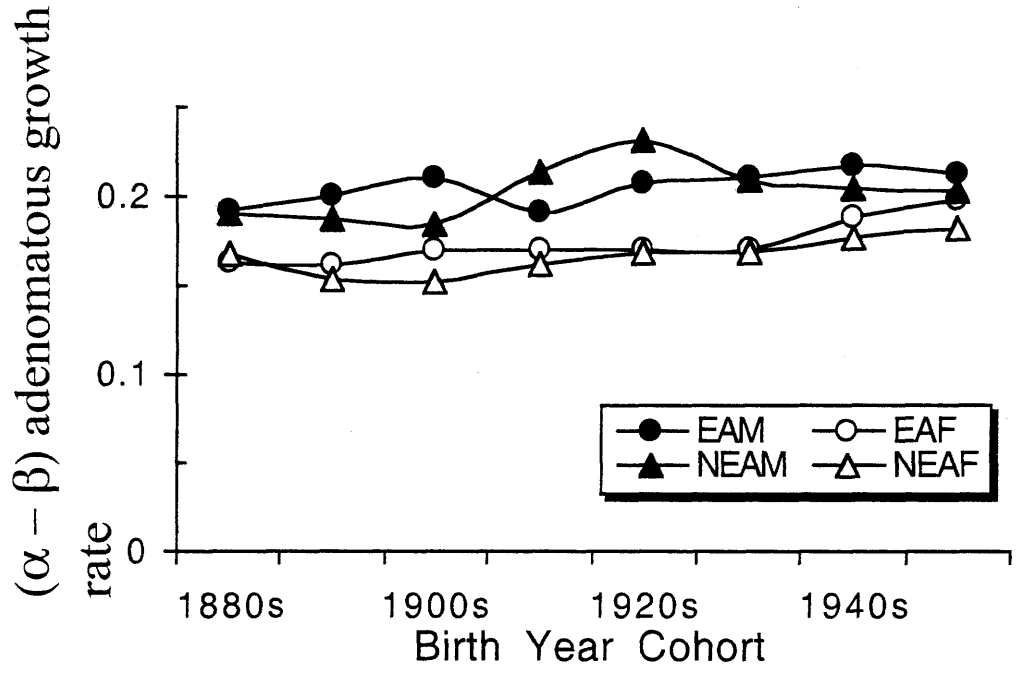

F_h - Fraction at risk
0.5
0.4
0.3
0.2
0.1
0
EAM
EAF
NEAM
NEAF
1840s
1860s
1880s
1900s
1920s
Birth Year Cohort
r_i – Initiation mutation rate (10^{-8})
10
8
6
4
2
0
EAM
EAF
NEAM
NEAF
1840s
1860s
1880s
1900s
1920s
Birth Year Cohort

Fig. 62: Historical trend in $\mathbf{r_A}$ and $(\alpha-\beta)$ for colon cancer

Historical trend in the calculated promotion mutation rate, $\mathbf{r_A}$, and the calculated growth rate of a precancerous lesion, $(\alpha-\beta)$.


r_A - Promotion mutation rate (10^{-8})
30
20
10
0
EAM
EAF
NEAM
NEAF
1840s
1860s
1880s
1900s
1920s
Birth Year Cohort


(α – β) adenomatous growth rate
0.2
0.1
0
EAM
EAF
NEAM
NEAF
1880s
1900s
1920s
1940s
Birth Year Cohort

Similarly, $(\alpha-\beta)$ could not be ascertained for birth years prior to 1880 since data for the age interval 20-60 are needed for this purpose and data were available only from the reporting year of 1930 forward. As this value appears invariant for each population cohort, other parameters were estimated assuming that the adenomatous growth rate is a constant.

4.1.2 Calculated Parameters for the Case of ($n = 2$ and $m > 1$)

The hypothesis that $m=1$ is attractive because the observed value for the promotion mutation rate was similar to the rate of LOH in T-cell precursors (Grist et al, 1992). This would be consistent with the hypothesis that the necessary event for promotion was the loss of heterozygosity of any of an undefined set of second gatekeeper genes, such that the fraction at genetic risk could be defined as the fraction of the population heterozygous for at least one of the second gatekeeper genes.

However, this hypothesis is inconsistent with the undisputed fact that colon tumors display a very high fraction (on average 0.22) of LOH and LOI distributed over all chromosomes (Vogelstein et al, 1988, 1989; Garcia-Patiño et al, 1998; Resta et al, 1998; Uhrhammer et al, 1999; Ragnarsson et al, 1999, Cui et al, 1998). There are about $9 \times 63 = 567$ lineal cell divisions between a first adenoma and first carcinoma cell in colon cancer ($9 \sim$ division rate, 63 years $\sim\Delta_h$). The rate of LOH or LOI to achieve a fraction of 0.22 from events in adenomatous growth alone would thus be $0.22/567 = 3.9 \times 10^{-4}$ LOH or LOI events per adenoma cell division.

This estimate can be considered in terms of the geometric means of the promotional mutation rates for different values of m. Calculations are summarized in Table 3.

Table 3. Calculated geometric mean of promotion mutation rates , m=1-5.

Data are for European-American females born in the 1860s.

(unordered = mutations can occur in any order,

ordered = mutations must occur in one particular order)

m	Unordered r_A	Ordered r_A
1	1.2×10^{-7}	1.2×10^{-7}
2	2.5×10^{-5}	3.5×10^{-5}
3	1.4×10^{-4}	2.8×10^{-4}
4	3.3×10^{-4}	8.5×10^{-4}
5	5.4×10^{-4}	1.6×10^{-3}

For the case of a required order of the promotion mutations, values of m = 3 and 4 yield mutation rates bracketing the LOH/LOI rate of 3.9×10^{-4}. If order were not required, the values for m = 4 and 5 bracket this value. Such calculations can be of use in considering the number of LOH plus LOI events that might be required in tumor promotion, but such considerations should not lose sight of the fact that there is no present evidence that either LOH or LOI events are required in promotion.

4.1.3 Robustness of the Model

In order to calculate the parameters of the three-stage carcinogenesis model for several birth year cohorts, maximum likelihood techniques were avoided by making several approximations. These approximations recognizably might have led to an inadequate determination of the actual values.

Survival data were obtained by observations from a small, possibly unrepresentative population (Eisenberg et al, 1968; NCI Monograph No. 6, 1961; Survival Report Number 5,

1976; Ries et al, 1983; Beart et al, 1995; SEER, 1993, 1997, 1999). Table 4 shows the effects of a 10% error in the estimate of the relative survival on the calculated parameters for one cohort, European-American females born in the 1880s. It is clear by inspection that the population risk parameters, $\mathbf{F_h}$ and $\mathbf{f_h}$, as well as the initiation mutation rate, $\mathbf{r_i}$, would not be seriously affected by this range of errors. However, the terms for adenomatous growth rate and mutation during promotion are more sensitive to this type of error.

Additionally, the reported survival data excluded diagnoses of colon cancer first detected at autopsy. This primarily occurs in the elderly, and may lead to a larger error in the estimated survival in the elderly. Table 3 shows the effects of a 10% error in the estimate of survival for individuals older than 75. Again by inspection, the population risk parameters $\mathbf{F_h}$ and $\mathbf{f_h}$ would not be seriously affected by such an error.

Table 4. Percentage Change in Parameter Estimates Given Errors in Data Sets
(Data for European-American females born in the 1880s were used.)

	$\mathbf{F_h}$	$\mathbf{f_h}$	$\mathbf{r_i}$	$\mathbf{r_A}$	$(\alpha - \beta)$	
+10% error in S(h,t)	-0.9	+1.2	-1.5	-12.4	+4.0	
-10% error in S(h,t)	+0.1	-0.1	+1.8	+8.9	-4.0	
+10% error in S(h,75+)	-1.1	+1.5	+9.3	-19.8	0	
-10% error in S(h,75+)	+2.0	-2.6	-10.8	+29.1	0	
+10% error in slope	+11.1	-12.9	-4.7	0	0	
-10% error in slope	-5.5	+7.4	+2.6	0	0	
+10% error in Δ_h	-9.0	-2.8	+22.1	-48.8	0	
-10% error in Δ_h	+5.2	-11.2	-11.7	+95.7	0	
+10% error in A_h	+3.0	+6.0	+3.5	0	0	
-10% error in A_h	+0.8	-11.0	-5.5	0	0	
+5 years in t_{max}	+10.9	-12.6	-17.0	0	0	
-5 years in t_{max}	+1.4	-1.7	+17.2	0	0	
+10% error in $\alpha-\beta$	0	0	-4.7	-38.0	0	
-10% error in $\alpha-\beta$	0	0	+5.4	+95.7	0	
MLT model*	+3.9	+3.9	-0.3	-21.8	+2.1	+2.0

*Maximum Likelihood Technique

Robustness of the approach can also be tested for how much an error in any one of the several observations derived from the raw mortality data by inspection would affect estimates of the parameters (Table 4). They alert to the uncertainty of the estimate of the promotional mutation rate, $\mathbf{r_A}$, while indicating a general robustness with regard to estimates of all other derived parameters.

Of additional interest is that the solution that would have been derived by the maximum likelihood routine (Table 4) showed no significant difference from that given by the five equations five unknowns methodology (excluding $\mathbf{r_A}$ once again), giving us the confidence that the approximations used in this methodology are sufficient to define the parameters of the three-stage carcinogenesis model.

4.1.4 5 Equations 5 Unknowns Methodology vs. Exact Solution (Moolgavkar and Luebeck, 1992)

Always of concern when using approximations in mathematical modeling is that in doing so, the end results may deviate from the exact solution with statistical significance. To verify that the approximation to the linear rising part of the hazard function, $\mathbf{P_{OBS}(h,t)}$, is not significantly different from the exact solution, we plugged into our own model the results for colon cancer incidence rates in Britain as given by the exact solution model of Moolgavkar and Luebeck (1992). (The growth rate of precancerous lesions were reported as an exponential rate which can be converted to a doubling rate for our purposes by dividing the estimates of α and β by ln 2).

Figure 63 plots the results for both methods on the same plot. The error between our solution and the exact solution to the hazard function, $\mathbf{P_{OBS}(h,t)}$, is approximately less than

4% (Figure 64) giving us reason to believe that our own methodology is sufficient. However, as verified in Figure 63, the exact solution does indeed deviate from our solution, bending over, but only at extremely high ages.

The linear estimate, as per Nordling (1953) for the case of two initiation mutations, is the asymptote to the exact solution, with approximated slope $2\tau^2 \mathbf{r_i r_j} N_{max} (\alpha - \beta)/\alpha$. Herrero et al (2000) calculated the exact slope of the asymptote, considering the possibility, albeit small, that a cell in a very small precancerous lesion, undergoes promotion, and then that precancerous lesion becomes stochastically extinct. (Our approximation only allowed promotion in precancerous lesions that have already survived stochastic extinction.)

Promotion in one of the cells of this to-be extinct lesion would only be of significance if the rate of promotion were high. Results from Table 2 (Section 4.1.1) however suggest that rates for the case of m = 1 are low. If more than one promotion mutation were required, we recognize that one of these rare events must occur after each promotion mutation, as each mutation gives rise to a new colony of cells that could become stochastically extinct.

Probability that a precancerous lesion survives stochastic extinction

$$2\tau^2 r_i r_j N_{max} \frac{\alpha - \beta}{\alpha}$$

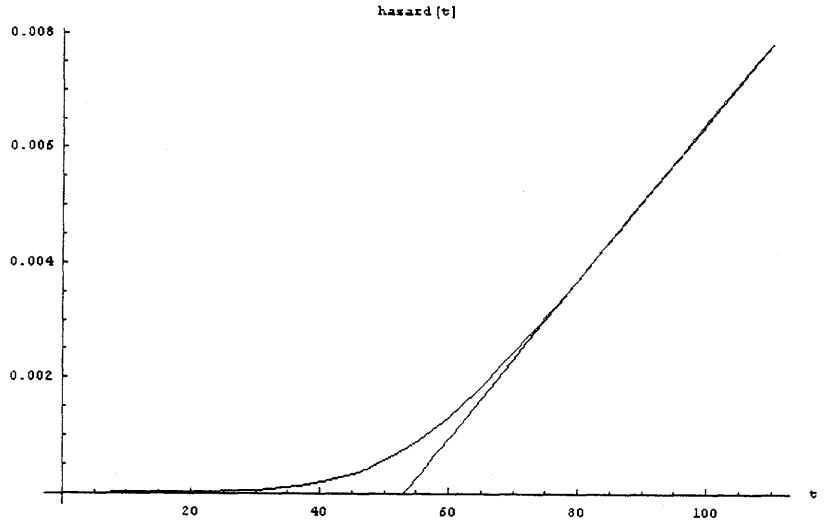
$$2\tau^2 r_i r_j N_{max} \left(1 + \frac{-(\alpha+\beta) + \sqrt{(\alpha+\beta)^2 - 4\alpha\beta(1 - r_A \frac{\alpha_c - \beta_c}{\alpha_c})}}{2(1 - r_A \frac{\alpha_c - \beta_c}{\alpha_c})\alpha} \right)$$

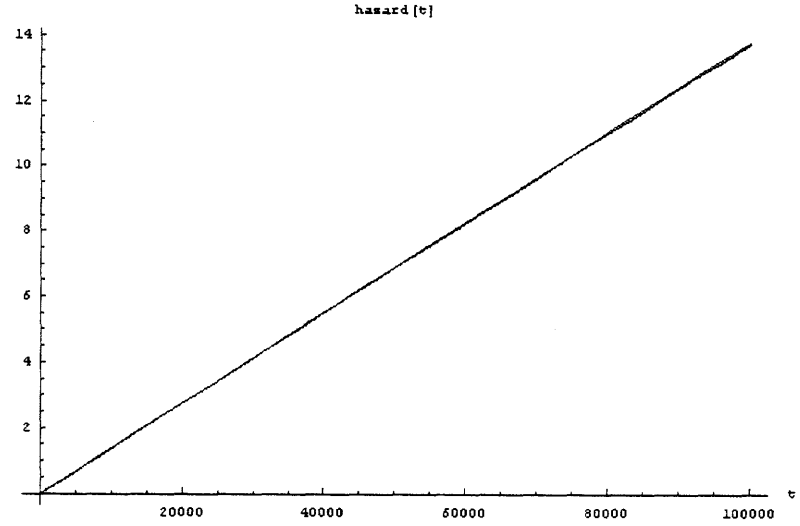
Probability that a cell in the precancerous lesion acquires the necessary promotion event (including a cell in a colony that later becomes extinct)

Fig. 63: $\mathbf{P_{OBS}(h,t)}$ - Comparison of exact solution to 5 Equations 5 Unknowns solution

Comparison of the exact solution of the hazard function $\mathbf{P_{OBS}(h,t)}$ Solution (Moolgavkar and Luebeck, 1992) to the approximation as used in the five equations five unknowns methodology (Sections 3.5.4 and 3.5.5).

We note that there is little deviation until ages $t > 100{,}000$ years of age given the estimated parameters of Moolgavkar and Luebeck (1992).


hazard[t]
0.008
0.006
0.004
0.002
20
40
60
80
100
t


hazard[t]
14
12
10
8
6
4
2
20000
40000
60000
80000
100000
t

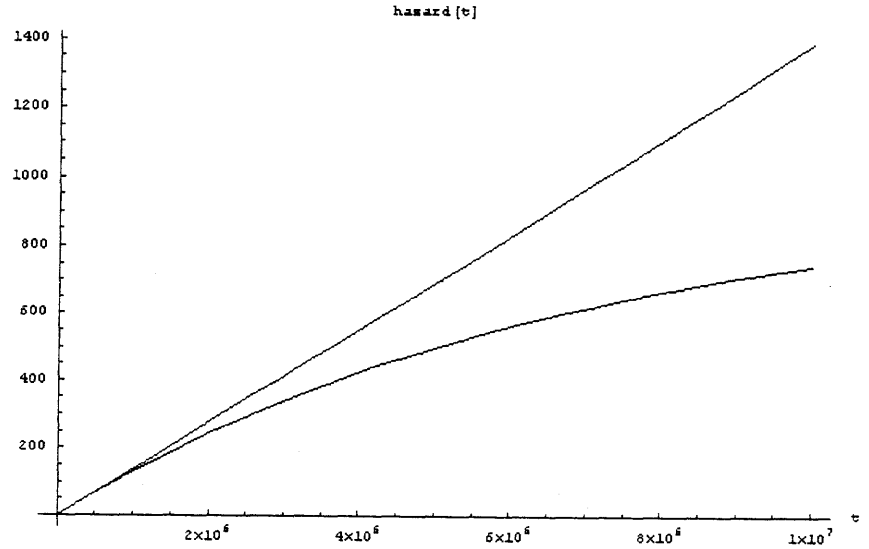
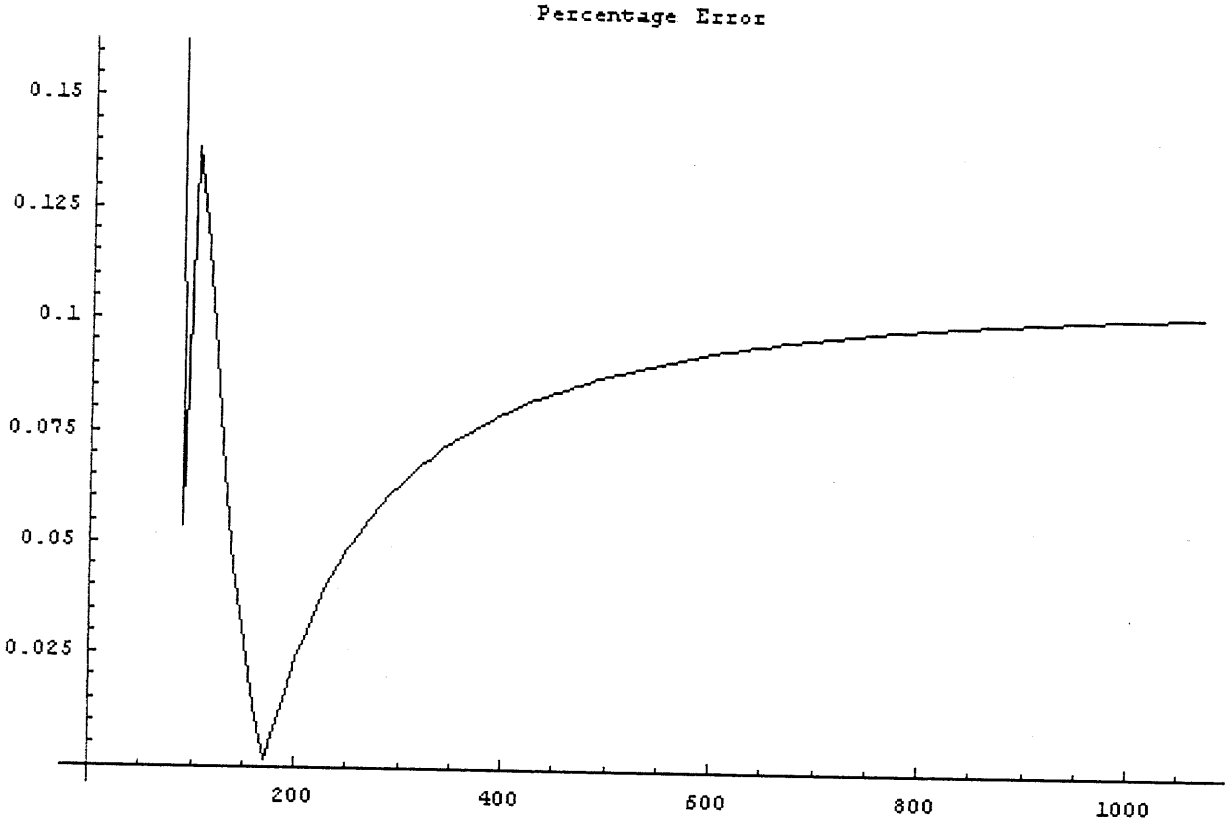
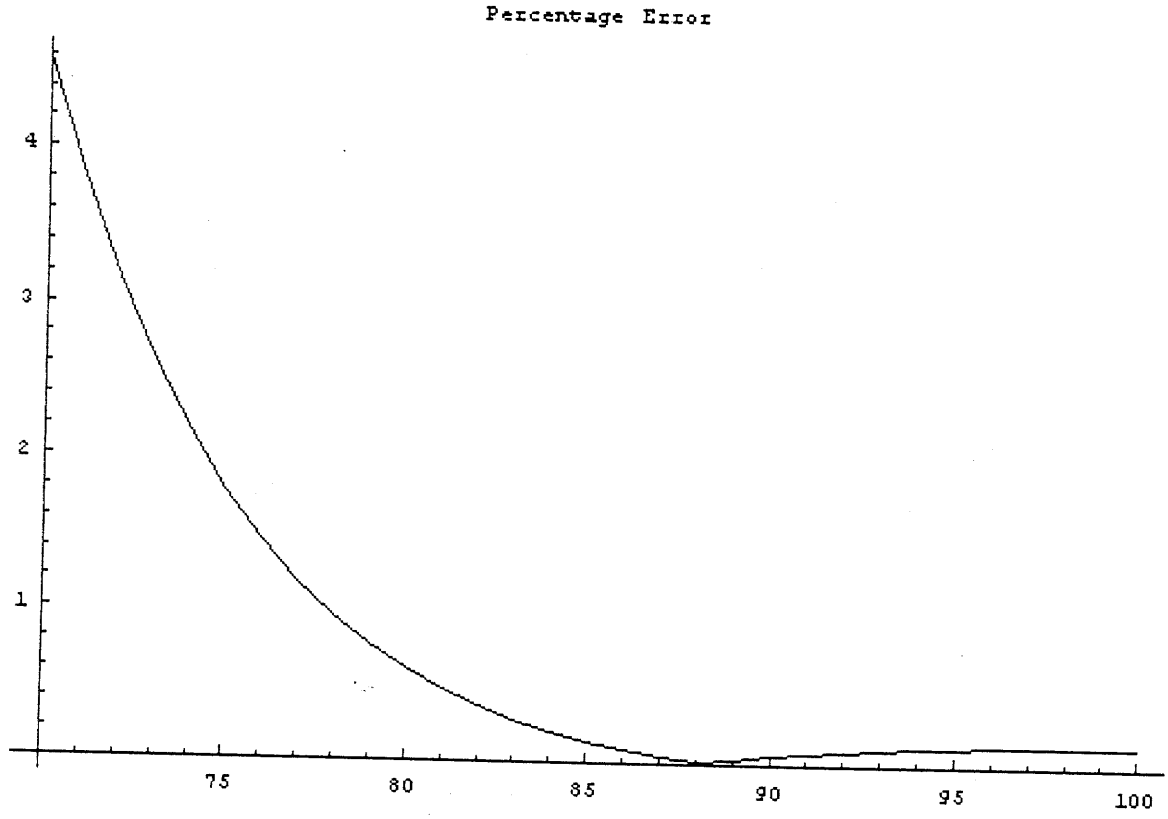

hazard[t]
1400
1200
1000
800
600
400
200
2x10^6
4x10^6
6x10^6
8x10^6
1x10^7
t

Fig. 64: Percentage error of the 5 Equations 5 Unknowns solution to $\mathbf{P_{OBS}(h,t)}$ versus exact solution

The ages of interest, $70 < t < 110$ which we use in the five equations five unknown methodology to calculate the slope, have an error $< 4\%$.

The turning over of the exact solution also does not appear to significantly occur before age 1000, as the error for ages 110 to 1000 is less than 0.1%.


Percentage Error
0.15
0.125
0.1
0.075
0.05
0.025
200
400
600
800
1000


Percentage Error
4
3
2
1
75
80
85
90
95
100

4.1.5 Turning Over of the Exact Solution to the Hazard Function, $\mathbf{P_{OBS}(h,t)}$

As shown in Figure 63, the exact solution of the hazard function does not continue to rise linearly at higher ages. At first glance, this would appear to not have been indicated by expectation. Nordling (1953) had argued that if the carcinogenesis process required two events, the hazard function, $\mathbf{P_{OBS}(h,t)}$, would rise linearly as a function of age, but recall that in the case that only one event were required, the mortality would be constant for all ages.

Suppose then that two initiation mutations were indeed required. As an individual ages, the percentage of cells that have not yet acquired the first initiation mutation decreases. In other words, in an individual who is still alive after a very long time, it is expected that that individual's organ would be primarily composed of cells containing the first of two initiation mutations. Therefore, the hazard function should rise linearly for the younger ages, as most of the cells have not yet mutated and follow a two-mutation process for initiation. Eventually, the hazard function is expected to bend and appear constant, since as the individual ages, eventually a majority of the cells would have the first mutation, and consequently the average cell would more approximately follow a one-mutation process for initiation.

It is imperative to assess the turning over of the hazard function, because in our model of the observed mortality curve, we have assumed that the curvature is due only to the existence of a subpopulation at risk that dies off at a faster rate than the remaining population. Any other form of curvature would lead to an underestimate of the fraction at risk.

Explored here is the rate at which the first initiation mutation must occur such that the turning over of the hazard function, $\mathbf{P_{OBS}(h,t)}$, is significant enough to create the condition that a large percentage of cells in a surviving individual are carrying the first initiation mutation during a normal lifetime.

Assumptions:

Both the first and second initiation mutations may occur in either a stem or transition cell.
No more initiation events can occur in a turnover unit where the stem cell has been initiated.

There are N_{stem} stem cells and N_{max} total cells.

There are N_{tu} cells per turnover unit.
X is the number of times the turnover unit has been renewed since birth
All cells except last layer of turnover unit divide once per renewal.
Divisions in turnover unit are synchronized.

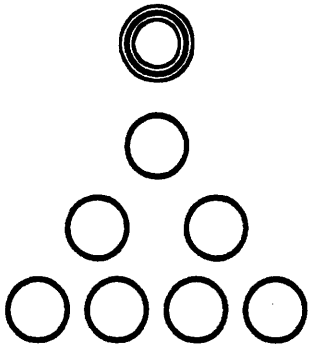
The turnover unit structure is:

0

1

10 11

100 101 110 111



where the stem cell is represented at the top, and each layer below represents the transition cells. Transition cells during turnover give rise to the 2 cells directly below. The stem cell gives rise to the stem cell and the first transition cell, while the last layer dies off during each turnover.

Numbers to the left correspond to the cell position. The reason for using binary notation is that the progenitor cell of a cell with number $N_1N_2N_3$ ...N_{Layer} is simply $N_1N_2N_3$ ...$N_{Layer-1}$. (The exception is that the stem cell (0) gives rise to (0) and (1), showing that this division is asymmetric).

We keep track of the state of the cell at each cell position in the turnover unit. The stages are:

State B(position #) – Normal (blank) cell

State I(position #) – Cell has first initiation mutation

State J(position #) – Cell is initiated

P_B, P_I, P_J represent the probabilities that the cell is in each corresponding state. There are going to be a total of (3 x N_{tu}) differential equations per turnover unit. The differential equations for ANY position are of the form:

$$\frac{dP_{B,position}}{dX} = (1-r_i)^2 P_{B,progenitor} - P_{B,position}$$

$$\frac{dP_{I,position}}{dX} = 2r_i(1-r_i)P_{B,progenitor} + (1-r_j)P_{I,progenitor} - P_{I,position}$$

$$\frac{dP_{J,position}}{dX} = r_i^2 P_{B,progenitor} + r_j P_{I,progenitor}$$

The exact hazard function for initiation in the turnover unit for the last renewal of the turnover unit is thus represented by:

$$\sum \frac{dP_{J,position}}{dX}$$

while our approximate hazard function for a turnover unit would have been:

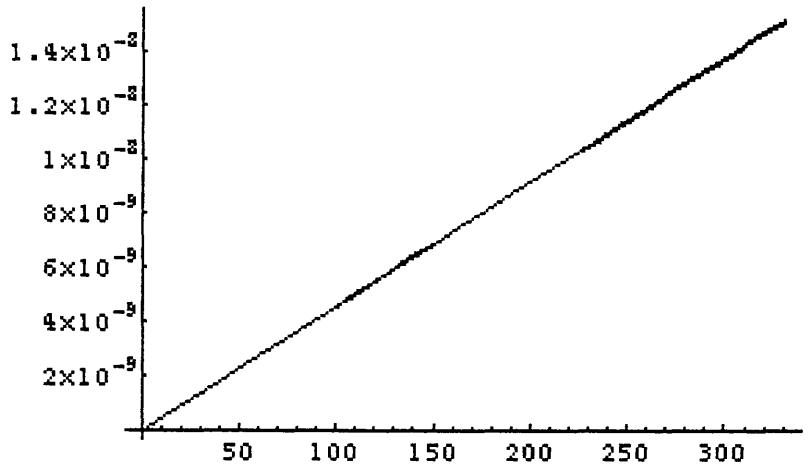
$$2\tau^2 r_i r_j N_{tu}$$

With the following conditions (from Results in Table 2, Section 4.1.1):

$r_i = 8 \times 10^{-8}$ $\qquad$ $r_j = 3r_i$

$\tau = 3$

and the sample case $N_{tu} = 256$ (real turnover unit size unknown), we can plot the resulting exact and approximate hazard functions for the turnover unit up to the ages of interest (110 years):



x-axis is divisions (plotted up to 110 τ)

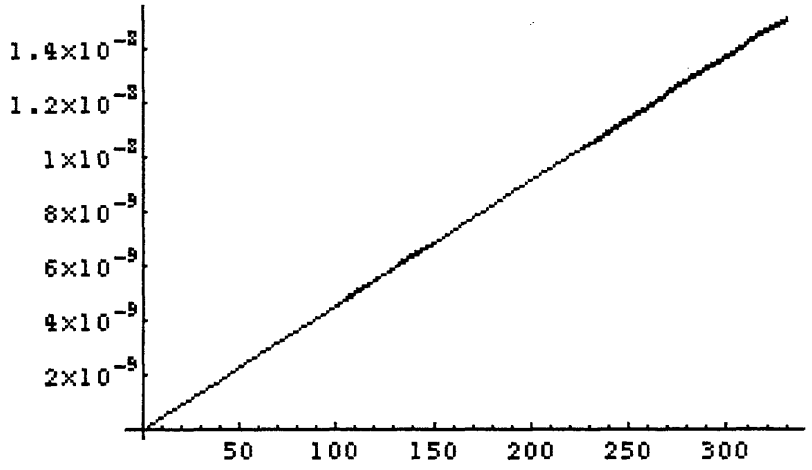
This demonstrates that for mutation rates estimated in Table 2 (Section 4.1.1) for colon cancer, the exact solution does not turn over significantly. Of course, the age at which the hazard functions turns over is going to depend on both the size of the turn over unit and the initiation mutation rates. We check for the following parameters:

Set 1	**Set 2**	**Set 3**	**Set 4**
$N_t = 256$	$N_t = 256$	$N_t = 256$	$N_t = 256$
$r_i = 1 \times 10^{-7}$	$r_i = 1 \times 10^{-6}$	$r_i = 1 \times 10^{-5}$	$r_i = 1 \times 10^{-4}$
$r_j = 3r_i$	$r_j = 3r_i$	$r_j = 3r_i$	$r_j = 3r_i$
$\tau = 3$	$\tau = 3$	$\tau = 3$	$\tau = 3$

We suspect that mutation rates will not be higher than those of set 3, since *in vivo* mutation rates of T-cells at the *hprt* locus are 100 fold lower. (Bigbee, 1998; Branda, 1993; Davies,

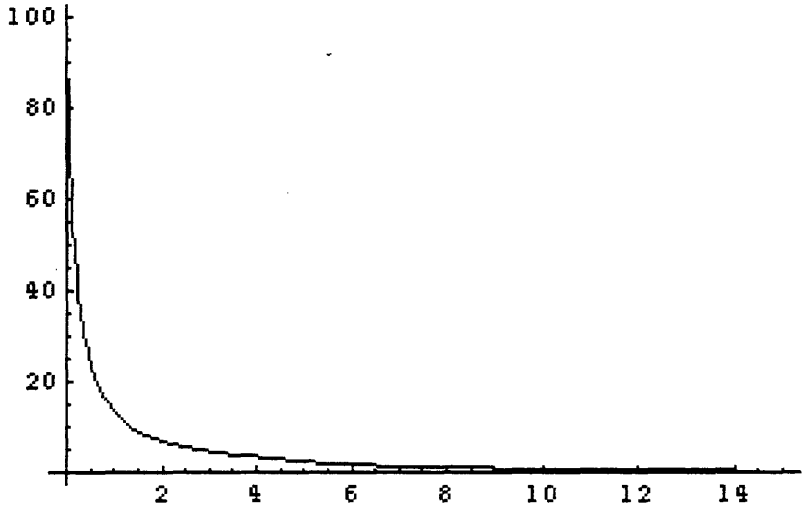
1992; Finette, 1994; Henderson, 1986; Hirai, 1995; Hou, 1995; Huttner, 1995; Liu, et al, 1997; McGinniss et al, 1990; Tates et al, 1991).

Set 1: Plot of exact and approximate hazard functions for the turnover unit on same graph



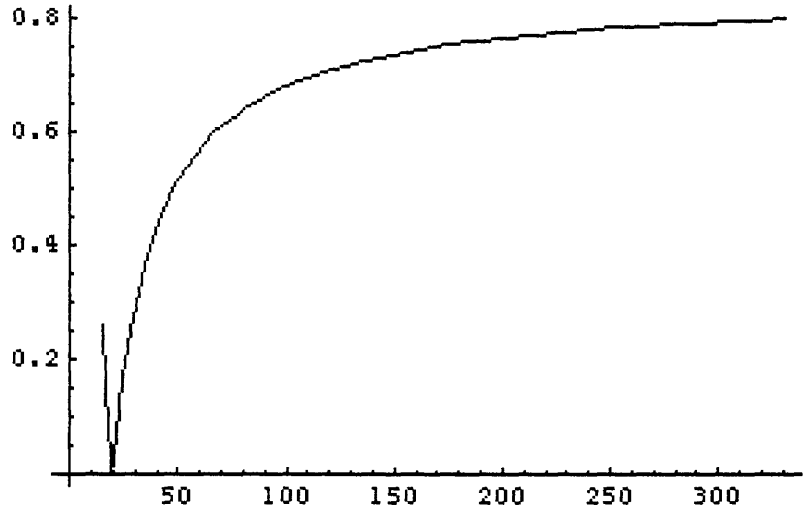
divisions (plotted up to 110 τ)

The percentage difference between the exact and approximated solutions is



number of divisions (first five years)

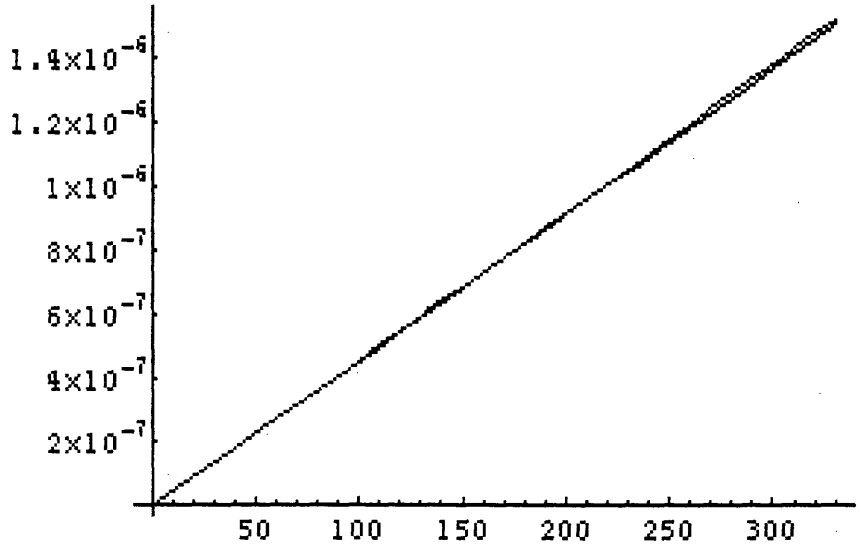
There is a large error during the first year, since the exact solution accounts for the delay between a stem cell acquiring the first initiation mutation and the time it takes to repopulate the turnover unit; the graph above is similar for all 4 sets.



error is < 1% for all ages between 5 and 110

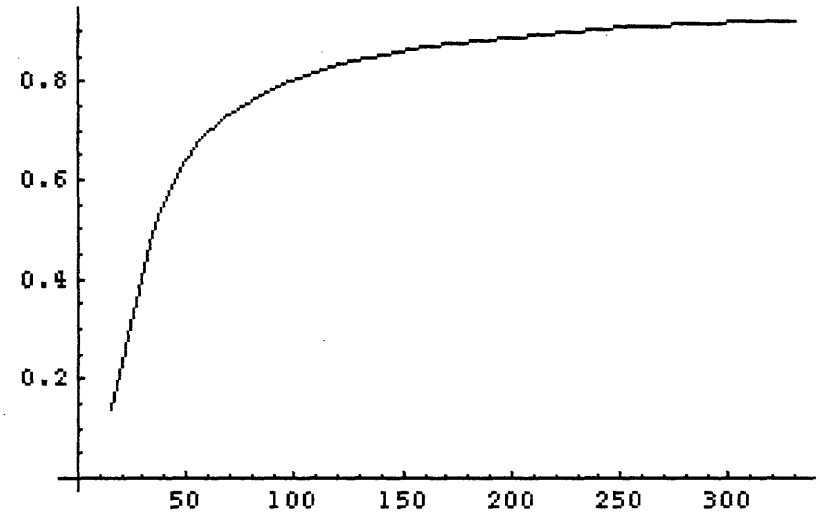
Set 2

Plot of exact and approximate hazard functions for the turnover unit on same graph



x-axis is divisions (plotted up to 110 τ)

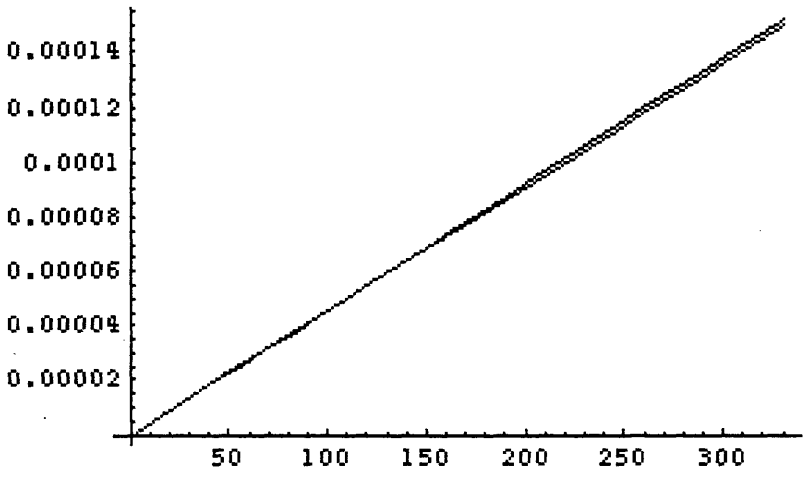
The absolute percentage difference between the exact and approximated solutions is



error is < 1% for all ages between 5 and 110

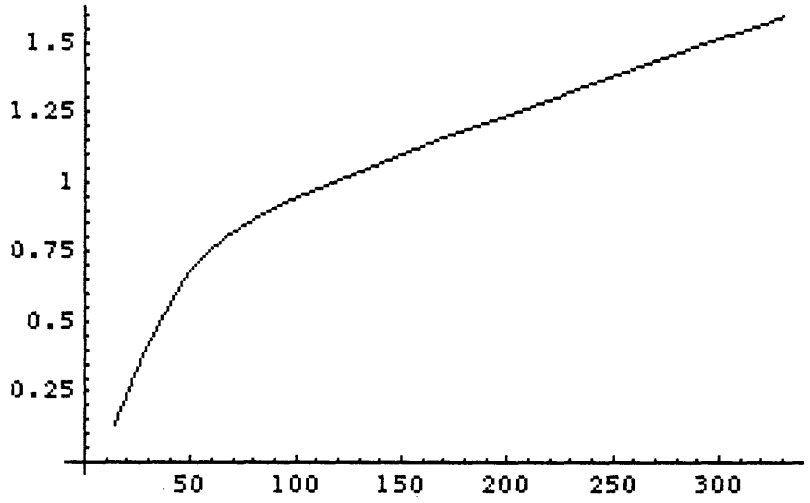
Set 3

Plot of exact and approximate hazard functions for the turnover unit on same graph



x-axis is divisions (plotted up to 110 τ)

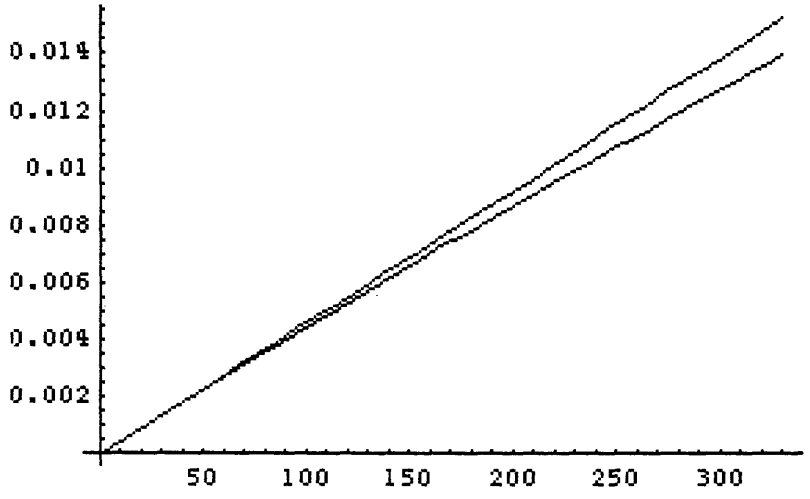
The absolute percentage difference between the exact and approximated solutions is



error is < 1.75% for all ages between 5 and 110

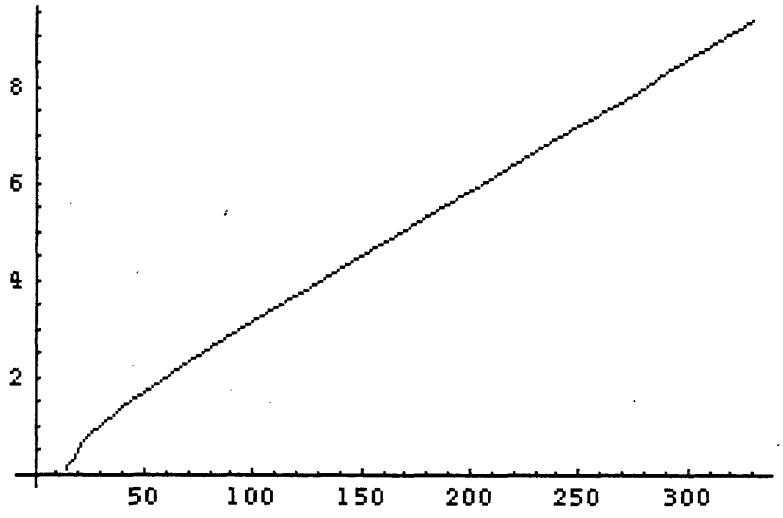
Set 4

Plot of exact and approximate hazard functions for the turnover unit on same graph



x-axis is divisions (plotted up to 110 τ)

The absolute percentage difference between the exact and approximated solutions is



error for all ages between 5 and 110

It would thus appear to be true that if mutation rates were less than 1×10^{-4} (Set 4) per cell division, the approximation should be more than sufficient. The curvature caused by the turning over from the depletion of stem cells not yet having acquired the first initiation mutation should therefore not affect our calculation of the fraction at risk given real physiological parameters.

4.2 Lung Cancer

4.2.1 Historical overview of lung cancer mortality in the American population.

To gain an overall historical perspective, one can first examine the lung cancer mortality rates for specific age groups (i.e. 60-64-year olds, Figure 65). There are clear differences expected from historical data regarding smoking prevalence between males and females. For the 60-64 year old age cohort, male lung cancer death rates were minimal for birth decades before 1860 but rose to a maximum value by the 1910-19 birth decade. For females,

lung cancer death rates were minimal and constant up to the birth decade of 1890 and then rose steadily not having reached a stable maximum by the birth decade of 1920. A clear historical difference of about 25 years demarked this gender-specific difference. There are no major historical differences between European American and African American populations. This similarity between these American ethnic subpopulations extends to colon cancer but not to all forms of cancer (Section 4.1.1).

The use of the entire U.S. population grouped into ten year birth year intervals assures a very high number of observations in each age and birth decade category. They are, with the exception of rates in children and the category 100+ years, which have little effect on calculations, generally smaller than the symbols in our figures. (i.e. 35 deaths among 0-4 year olds born in the 1980s, and 76 reported deaths among 100+ year olds born in the 1880s.) The complete numerical data set is available for review and further research. (See end of ABSTRACT for instructions).

4.2.2 Calculated parameters for the case of (n = 2 and m = 1) - Nonsmokers

Mortality rates among women born in the 1820s to the 1880s from lung cancer were minimal and constant prior to cigarette adoption by each successive cohort. (Figure 65) These earliest data create an age-specific lung cancer mortality function for the pre-cigarette smoking female birth year cohorts. (Figure 66) Comparing these results to independently reported lung cancer mortality rates from the American Cancer Society's second cancer prevention study (CPS-II) of a much smaller group of one million Americans during the 1980s reveal no significant differences. (Peto et al, 1988, 1992). (Studies by Enstrom et al (1980), Kahn (1966) and Doll (1968) are also consistent with the data of Figure 66.

Fig. 65: Historical trend in lung cancer mortality for 60-64 year olds

A clear historical difference of about 25 years in lung cancer mortality demarks the gender-specific difference of smoking prevalence.

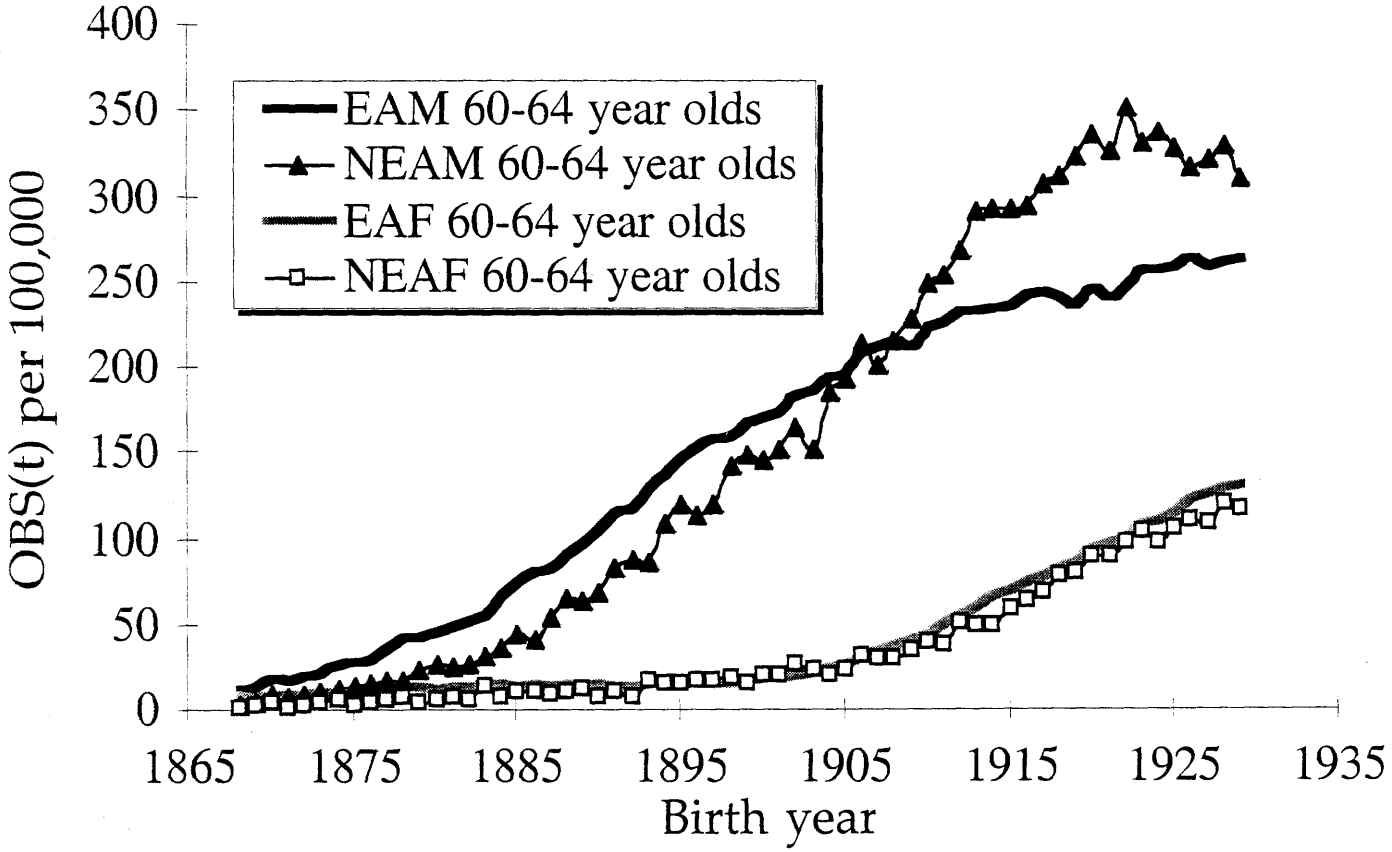
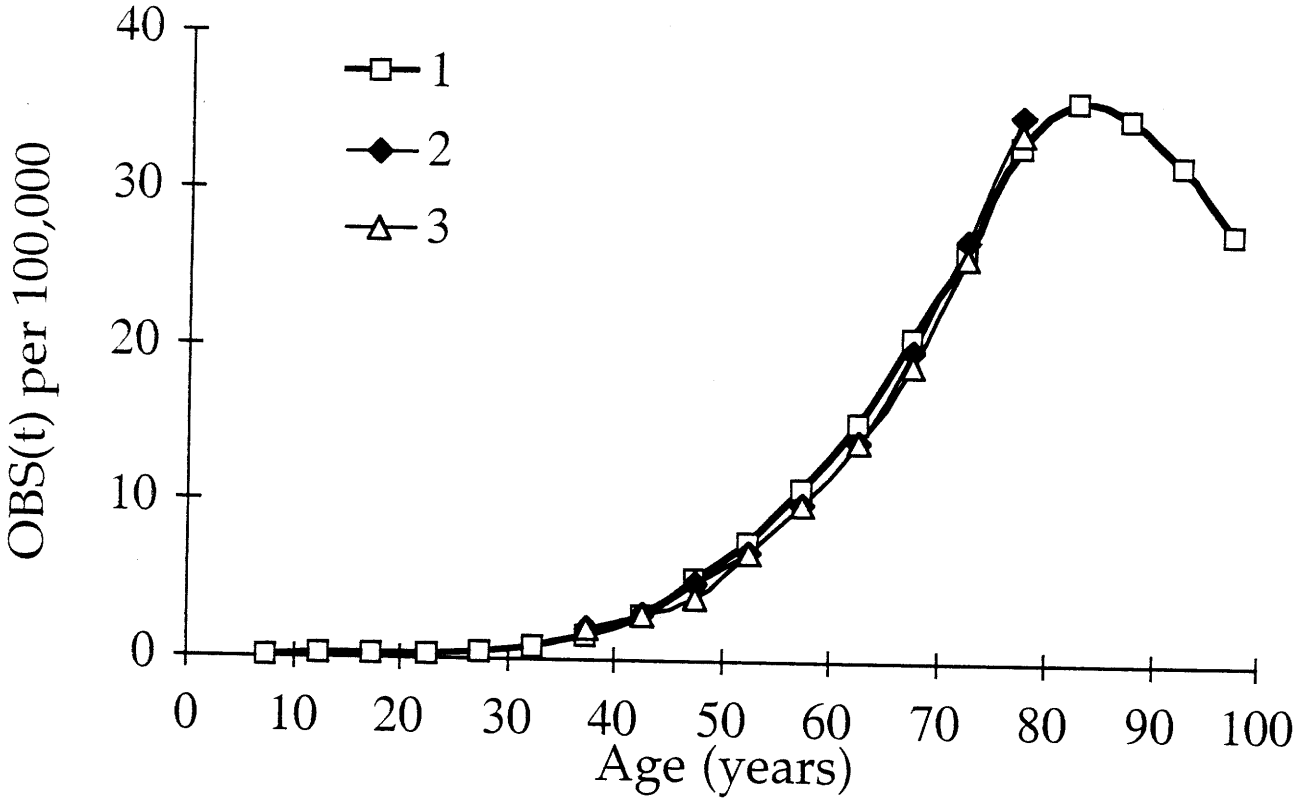

400
350
300
250
200
150
100
50
0
OBS(t) per 100,000
EAM 60-64 year olds
NEAM 60-64 year olds
EAF 60-64 year olds
NEAF 60-64 year olds
1865
1875
1885
1895
1905
1915
1925
1935
Birth year

Fig. 66: Lung cancer age-specific birthyear specific mortality rates for nonsmokers:

1) European American female nonsmokers (reconstructed from mortality rates for birth decade cohorts prior to the 1880s)
2) American Cancer Society's second cancer prevention study - female nonsmokers (Peto 1988, 1992)
3) American Cancer Society's second cancer prevention study - male nonsmokers. (Peto 1998, 1992)

Data of (2) and (3) were smoothed by the authors; the raw data show a greater scatter as would be expected for the smaller cohorts than that comprised in (1).

Peto et al (1988, 1992) also have an estimate for 85+, higher than our estimate of the mortality using female nonsmokers born prior to the 1880s. However these results are not inconsistent, since Peto et al's studies are more recent and less likely to be confounded by underreporting. The effect of a decreased chance in underreporting on a mortality/incidence curve is predicted by the three-stage carcinogenesis model to result in an increase in $\mathbf{f_h}$, thereby increasing observed mortality rates only among the elderly. (as shown in Figure 52)


40
30
20
10
0
OBS(t) per 100,000
1
2
3
0
10
20
30
40
50
60
70
80
90
100
Age (years)

The reconstruction used the data for all females in the U.S. population dying of lung cancer born between approximately 1820 and 1890. The CPS-II populations of were lifetime nonsmokers for which smoking status of the deceased was individually determined were born between approximately 1900 and 1970. The degree of agreement between the reconstruction of the nonsmoker age-specific lung cancer death rates and those of CPS-II and similar studies is obvious by inspection. This agreement gives confidence to the estimates of mortality rates at ages greater than 75 years, which are essential for calculation of risk parameters in the quantitative carcinogenesis model.

An equivalent attempt for males is confounded by lack of data for birth years preceding the general use of cigarettes by a significant fraction of American males. The estimates from the American Cancer Society's second cancer prevention study (Peto et al 1988, 1992) however suggest that lung cancer mortality rates among nonsmoking males were not different from those among nonsmoking females. The female data set is therefore used to make first order estimates of the parameters of lung carcinogenesis among nonsmoking men and women.

From the cross-sectional reconstruction of the age-specific lung cancer mortality for European American females born prior to the 1880s, the estimated value of ($\mathbf{F_h} = \mathbf{F_{NS}}$) is 0.1, and of $\mathbf{f_h}$ is 0.15.

The fraction of these obligatory nonsmokers to be at lifetime or essential risk of lung cancer is about 10%. (The actual fraction of persons dying of lung cancer in these nonsmoking birth cohorts was actually, as expected, much lower at about 0.7%.) Deaths from lung cancer represented only about 15% of all deaths caused due to competing forms of death sharing the identical inherited and environmental essential risk factors for lung cancer in nonsmokers. The age-specific mortality data for mid 19th century females in the United States and reported

observations of these same rates among American nonsmokers of the present era as shown in Figure 66 are remarkably similar. It appears that the fraction of nonsmokers at lifetime risk of lung cancer has shown little if any historical variation. This is similar to the finding with regard to an historically invariant fraction of the United States population at lifetime risk for colon cancer (Section 4.1.1)

Assuming *pro tempore* that half of all lung cancer cases occur in the upper bronchi, permits an estimate for the initiation parameter, $2\ \tau^2\ \mathbf{r}_i\ \mathbf{r}_j$, of 3.4×10^{-11} in the tracheal bronchial and 2.9×10^{-13} for the peripheral bronchiolar region of the lung. As may be noted in Table 4, varying the fraction of tumors between these regions modifies these calculations to but a small extent.

Figure 66 permits a straightforward estimate that $(\alpha - \beta) = 0.17$ doublings per year in hypothetically exponentially growing preneoplastic colonies in the lungs of nonsmokers at lifetime risk of lung cancer. This value applies to all tumors of the lung regardless of anatomical location and is approximately the same as previously calculated for the human colon. (Section 4.1.1)

Given the values of the population and other physiological parameters it is possible to estimate rate values for any theoretical biological model of promotion in nonsmokers. For the case of a single required promotional event, m=1, one estimates a rate of $\mathbf{r}_A = 2.8 \times 10^{-7}$ events per cell division, a value again similar to that estimated for human colon. (Section 4.1.1)

4.2.3 Calculated parameters for the case of (n = 2 and m = 1) – Smokers

Initial analysis is done on the age-specific mortality data for a birth decade cohort for which the predominant fraction of lung cancer deaths were among smokers and for which

mortality data exists from early to late adulthood. The cohort of European American males born 1910-1919 had a maximum smoking prevalence of 0.69. This cohort adopted the cigarette habit at an average age of 18 and provided sufficient mortality data for analysis of all risk parameters. (Harris, 1983) Using the estimations of population and physiological parameters for this cohort as a starting point, analyses to all other cohorts was made, subject to certain idiosyncratic restrictions due to inadequate data noted in the legend of Table 5. Key initial approximations are that all members of the cohort began smoking at the average age for the cohort and that the fraction of smokers for the cohort was the maximum fraction reported by the cohort in health questionnaires as summarized by Harris (1983) and noted in Figure 33.

For the cohort of European American males of the 1910s, one can estimate the value of ($\mathbf{F_h} = \mathbf{F_{NS}} \cup \mathbf{F_S}$) to be 0.71, and of $\mathbf{f_h}$ to be 0.17. This means that the estimated fraction of this cohort to be at lifetime or essential risk of lung cancer is about equal to the fraction of smokers in the cohort. Deaths from lung cancer represented only about 17% of all deaths sharing the identical inherited and environmental essential risk factors for lung cancer. This lifetime risk factor showed a clear historical increase among both males and females while the parameter accounting for competing risks remained historically invariant. (Table 5)

Estimations of the rates of mutations putatively required for initiation are critically dependent on the assumptions made about the number of required events, n, and the number of cells, N_a, which can give rise to preneoplastic colonies given the necessary event(s). For smokers, the initiation parameter, $2\,\tau^2\,\mathbf{r_i}\,\mathbf{r_j}$, was found to be about 3.1×10^{-11} for the tracheal-bronchial and 2.7×10^{-13} for the peripheral bronchiolar region. Varying assumptions regarding anatomical distribution of tumors in the lung had but little effect on the primary observation that the initiation parameter is much greater in the tracheal bronchial than the peripheral

bronchiolar region. (Table 5) (A Connecticut cohort of men born in the 1930s was reported to have somewhat more than 50% of all lung tumors in the upper bronchial tract so use of 50% of tumors in the upper tract is consistent with these observations). (Zheng et al, 1994; Thun et al, 1997) All of these estimates were historically invariant and remarkably similar to estimates for nonsmokers at lifetime risk of lung cancer. The value of the initiation parameter estimate for the peripheral bronchiolar region is almost exactly equal to the previous estimate of initiation rates in the colon. (Section 4.1.1)

Section 3.5.5.4 (Equation 30) demonstrates the calculation of the growth rate of precancerous lesions directly from cancer mortality data without need for maximum likelihood routines. For the colon, the growth rate of adenomas was constant throughout all the reported birthyear cohorts. (Figure 67) Contrarily, this methodology applied to the lung cancer mortality data set reveals that the calculated growth rate of precancerous lesions increases among birthyear cohorts with higher smoking prevalence. (Figure 67)

To explain why the growth rate of precancerous lesions in smokers appears to increase historically, one must first take into account the effect of having two potentially independent groups at risk for lung cancer. The growth rates are calculated by simply taking the $\log_2$ of the derivative of the observed mortality data, giving:

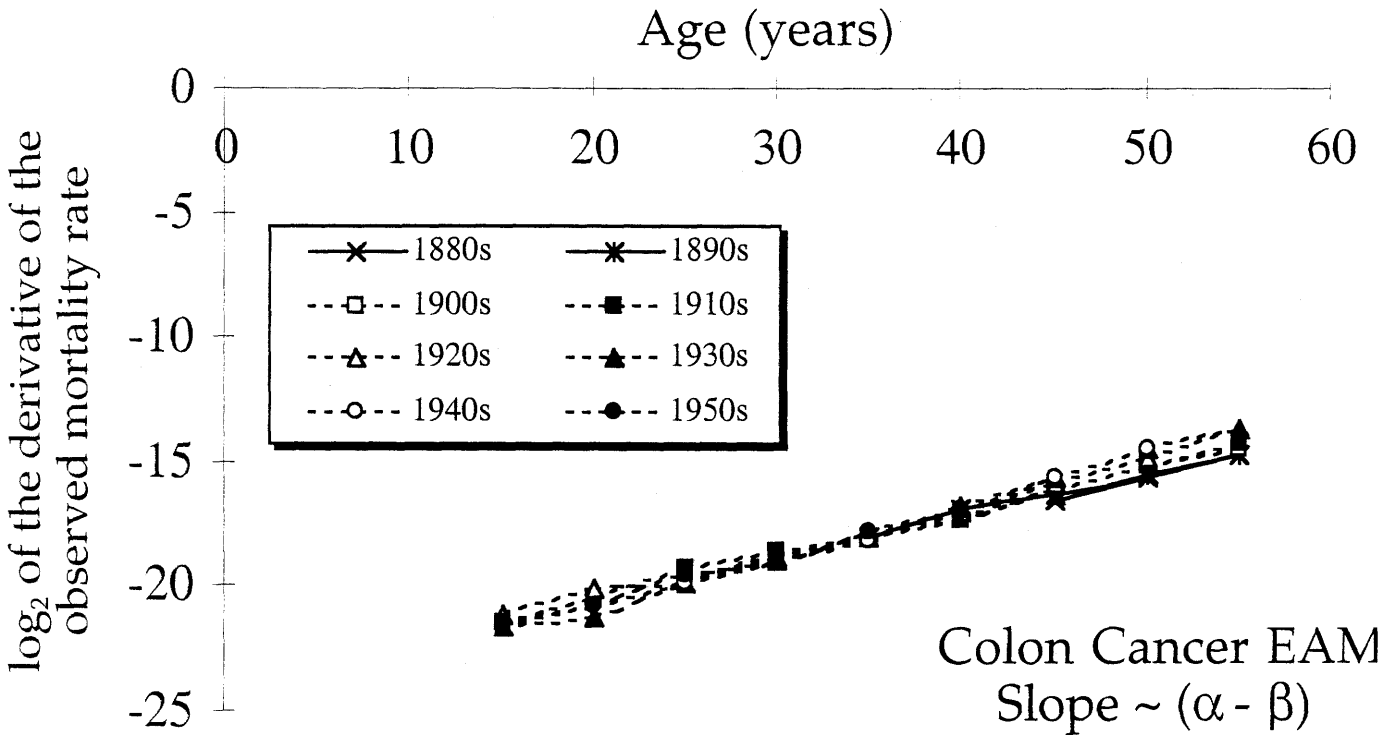
$$\log_2\left(\frac{\partial OBS^*(h,t)}{\partial t}\right) \sim \log_2(C_1\, 2^{(\alpha-\beta)\,t}) \sim \log_2(C_1) + (\alpha - \beta)t$$

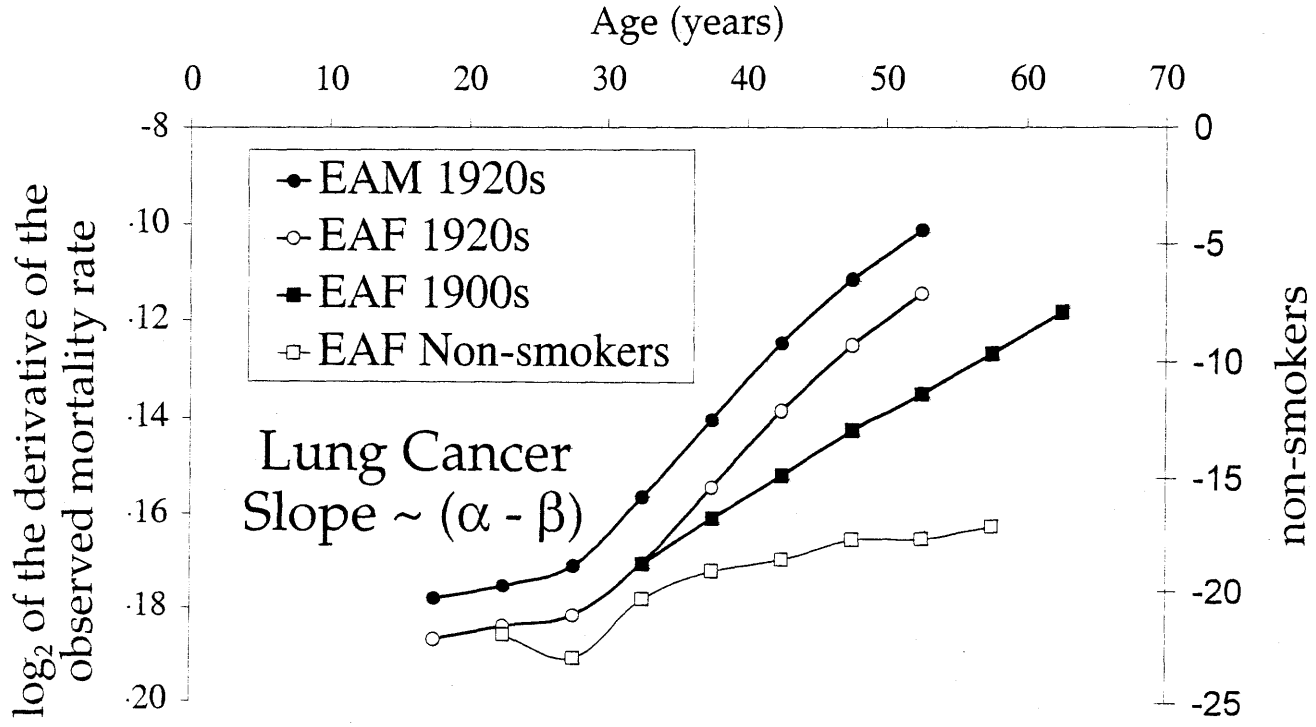
where C_1 is a constant associated with the values for the fraction at risk, mutation rates, and the adenomatous growth rate. This gives a line with slope $(\alpha - \beta)$.

Fig. 67: Graphical estimation of the growth rate of precancerous lesions of the colon and lung

Calculated as the slope of the $\log_2$ of the derivative of OBS*(h,t), the mortality rate adjusted for survival and underreporting. (Equation 4, Section 3.2.6)

Smoking prevalence: EAF 1900s ~ 23%, EAF 1920s ~ 42%, EAM 1920s ~ 69%


Age (years)
0
10
20
30
40
50
60
0
-5
-10
-15
-20
-25
log2 of the derivative of the observed mortality rate
1880s
1890s
1900s
1910s
1920s
1930s
1940s
1950s
Colon Cancer EAM
Slope ~ (α - β)


Age (years)
0
10
20
30
40
50
60
70
-8
-10
-12
-14
-16
-18
-20
0
-5
-10
-15
-20
-25
log2 of the derivative of the observed mortality rate
non-smokers
EAM 1920s
EAF 1920s
EAF 1900s
EAF Non-smokers
Lung Cancer
Slope ~ (α - β)

However, this equation is only true if there were only one group at risk. For lung cancer, the population consists of the union of three groups at risk: (a) individuals who do not smoke, but are still at risk for lung cancer from independent risk factors, (b) individuals who smoke and are placed at risk only for cigarette-induced lung cancer, and (c) individuals who smoke and are both at risk for cigarette-induced lung cancer and lung cancer due to independent risk factors. Accounting for all groups at risk would affect the calculated slope of the equation above. As a simplification, the effect of two groups is shown to be:

$$\log_2\left(\frac{d\mathrm{OBS}^*(h,t)}{dt}\right) \sim \log_2\left(C_1\, 2^{(\alpha-\beta)_1 t} + C_2\, 2^{(\alpha-\beta)_2 (t-(\Delta_2-\Delta_1))}\right)$$

where $(\Delta_2 - \Delta_1)$ represents the difference in the expected time until promotion for the two groups at risk (i.e. smoker's Δ_h vs non-smokers' Δ_h).

This function still behaves approximately linearly, with the overall slope now dependent on the adenomatous growth rates of each group at risk $(\alpha - \beta)_{1,2}$ and the difference in the expected time until promotion $(\Delta_2 - \Delta_1)$, as demonstrated in Figure 68 using values of $(\alpha - \beta)$ of 0.17 and 0.32 and setting for the example, $C_1 = C_2 = 1$. If the time until promotion for the two cohorts were different, then the estimated value for the precancerous lesion growth rate $(\alpha - \beta)$ for a mixed cohort using the slope of the $\log_2$ of the derivative of OBS*(h,t), would be an approximation of the growth rate for whichever group at risk had the shorter time until promotion, Δ_h.

In practice, this means that for a birthyear cohort such as European American Females born in the 1880s which began smoking at age 31 (Harris, 1983), since the lung cancer

mortality peak associated with cigarette smoking occurs later than the lung cancer mortality peak associated with other causes (Figure 27), the calculated value of (α - β) for this cohort is no longer representative of the growth rate of precancerous lesions of smokers in that cohort, but rather that of the nonsmokers in that cohort..

As the age of initiation of cigarette smoking decreases as a historical function of the birthyear cohorts, then our evaluation of (α - β) would begin to more accurately estimate the growth rate of precancerous lesions of smokers. The increasing trend in the values for the growth rates suggested by Figure 67 is therefore representative of the fact that individuals from each cohort are starting to smoke at an earlier age, as well as the fact that the fraction of smokers to non-smokers has increased historically.

The growth rates of precancerous lesions of smokers is therefore best represented by the estimates from the later birthyear cohorts, ~(0.31-0.32).

To calculate promotion mutation rates in smokers, one must further the age when the average smoker in a cohort began to smoke. As argued by Cook et al (1969), the x-intercept of the linear portion of the mortality curves, Δ, is equal to the time between initiation and promotion, Δ_h, plus the average age at which exposure to any necessary environmental risk factor(s) commenced. Correcting for the age at which individuals began to smoke, one can estimate the true value of Δ_h and $\mathbf{r}_A$ among smokers. For the case of a single required promotional event, $\mathbf{m}$=1, one can estimate a promotion mutation rate $\mathbf{r}_A$ of 3.7×10^{-7} events per cell division. The value of the promotion mutation rate estimate is similar to that calculated above for nonsmokers as well as to the previous estimate in the colon. (Section 4.1.1)

Fig. 68: Estimation of $(\alpha - \beta)$ in a cohort with two subpopulations at risk for death

Effect of the difference in the expected time until promotion between two groups at risk, $(\Delta_2 - \Delta_1)$, on the estimated value of $(\alpha - \beta)$ for a mixed cohort consisting of these two groups at risk. In this example, the two cohorts are assumed to be identical, except for the growth rates of their precancerous lesions, $(\alpha - \beta)$. Values of 0.19 and 0.32 are used.

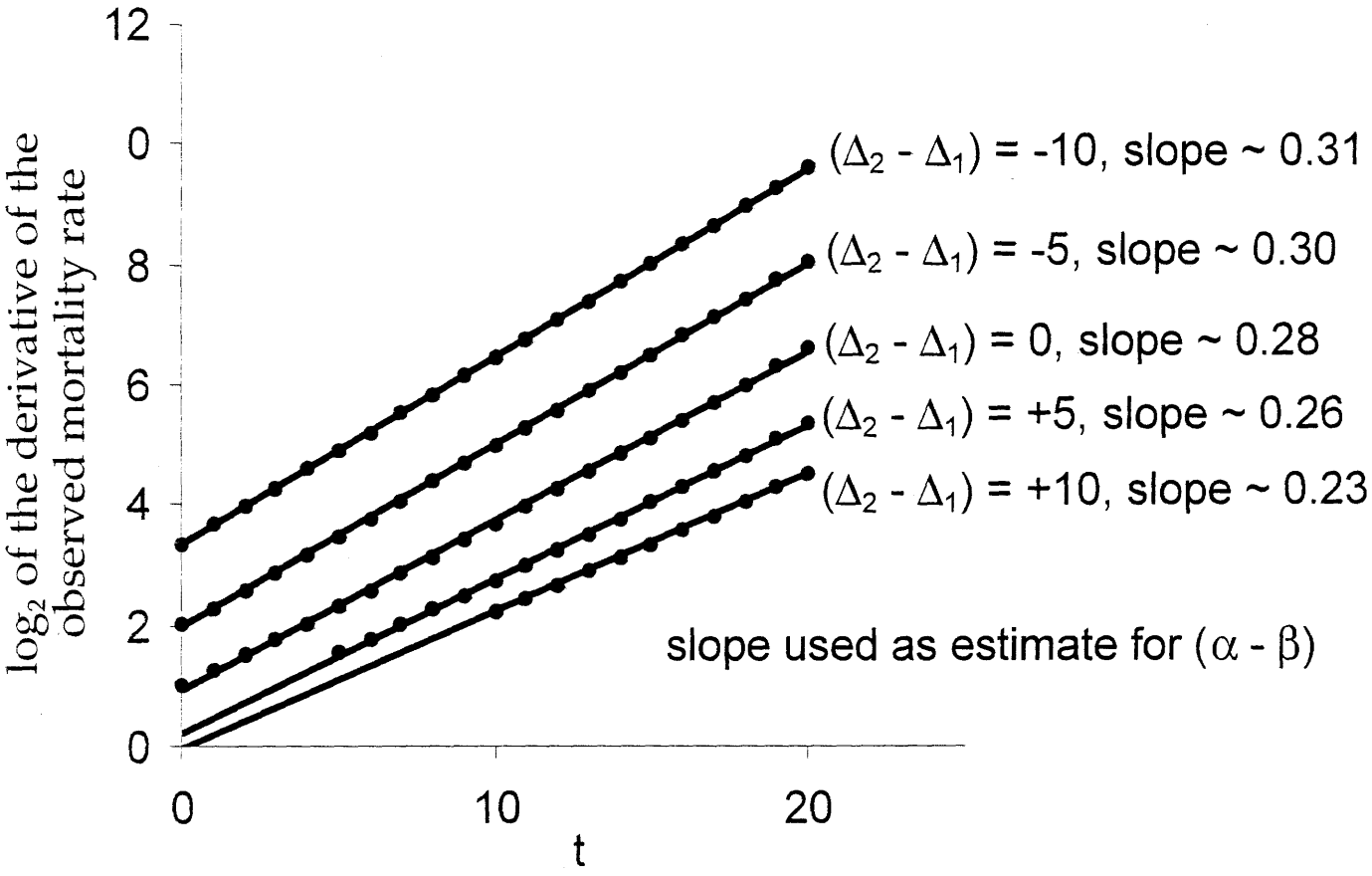

12
0
8
6
4
2
0
0
10
20
t
log2 of the derivative of the observed mortality rate
(Δ2 - Δ1) = -10, slope ~ 0.31
(Δ2 - Δ1) = -5, slope ~ 0.30
(Δ2 - Δ1) = 0, slope ~ 0.28
(Δ2 - Δ1) = +5, slope ~ 0.26
(Δ2 - Δ1) = +10, slope ~ 0.23
slope used as estimate for (α - β)

Table 5. Summary of calculated parameters for lung cancer.

	Smoking prevalence	Lifetime risk fraction	Competing Death Fraction	Initiation Rate	Initiation Rate	Promotion Rate	Preneoplastic Growth Rate
	$E_{h,cigarette}$	F_h	f_h	$2\tau^2 r_i r_j$ (tracheal bronchial) $N_{max} \sim 2.4 \times 10^8$	$2\tau^2 r_i r_j$ (peripheral bronchiolar) $N_{max} \sim 2.8 \times 10^{10}$	r_A	$(\alpha - \beta)$
FEMALES							
Nonsmokers	0	0.10	0.15	3.4×10^{-11}	2.9×10^{-13}	2.8×10^{-7}	0.17
1880s	0.03	0.17	0.13	3.8×10^{-11}	3.3×10^{-13}	1.8×10^{-8}†	0.31
1890s	0.10	0.24	0.13	2.1×10^{-11}	1.8×10^{-13}	1.5×10^{-7}	0.31
1900s	0.23	0.31	0.11	2.5×10^{-11}	2.1×10^{-13}	4.2×10^{-7}	0.31
1910s	0.37	0.41	0.13	2.8×10^{-11}	2.4×10^{-13}	4.2×10^{-7}	0.31
1920s	0.42	0.44	0.13	2.7×10^{-11}	2.3×10^{-13}	7.5×10^{-7}	0.31
MALES							
1870s	-	0.27	0.14	2.5×10^{-11}	2.2×10^{-13}	††	0.32
1880s	0.34	0.37	0.28	2.4×10^{-11}	2.1×10^{-13}	3.0×10^{-7}	0.32
1890s	0.49	0.53	0.23	3.0×10^{-11}	2.6×10^{-13}	2.6×10^{-7}	0.32
1900s	0.61	0.61	0.22	3.6×10^{-11}	3.1×10^{-13}	2.7×10^{-7}	0.32
1910s	0.68	0.71	0.17	3.1×10^{-11}	2.7×10^{-13}	3.1×10^{-7}	0.32
1920s	0.69	0.70	0.17	3.2×10^{-11}	2.7×10^{-13}	3.7×10^{-7}	0.32
MALES 1920s							
25% upper bronchi	0.69	0.70	0.17	1.6×10^{-11}	4.1×10^{-13}	3.7×10^{-7}	0.32
75% upper bronchi	0.69	0.70	0.17	4.8×10^{-11}	1.4×10^{-13}	3.7×10^{-7}	0.32
		F_{colon}			$2\tau^2 r_i r_j$ $N_{max} \sim 8.5 \times 10^{10}$		
Females	-	0.40	0.16	-	2.7×10^{-13}	2.0×10^{-7}	0.17
Males	-	0.41	0.21	-	2.3×10^{-13}	8.0×10^{-8}	0.20

† Lung cancer deaths in this cohort were predominantly due to nonsmoking risk factors (only 2-3% smoked) making evaluation of Δ_h for smokers particularly suspect, further affecting evaluation of the promotion mutation rate for this cohort.

†† Age of smoking initiation unknown.

4.2.4 Calculated Parameters for the Case of (n = 2 and m > 1)

If the required events for promotion required loss of heterozygosity or imprinting, the hypothesis that m=1 would be inconsistent with the undisputed fact that lung tumors display a very high fraction (on average 0.6 for either allele) of LOH and LOI distributed over all chromosomes studied (Wistuba et al 1997, 1999). The rate of LOH or LOI to achieve a fraction of 0.6 from events in precancerous growth alone would be $0.6/(2x406.4) = 7.4 \times 10^{-4}$ LOH or LOI events per cell division.

This estimate can be considered in terms of the geometric means of the promotional mutation rates for different values of 'm'. Calculations are summarized in Table 6.

Table 6. Calculated promotional event rates for m=1 to 6 necessary events in smokers independent of order of mutations

	r_A of smokers		r_A of smokers
m = 1	3.0×10^{-7}	m = 4	4.9×10^{-4}
2	4.4×10^{-5}	5	7.7×10^{-4}
3	2.2×10^{-4}	6	1.0×10^{-3}

4.2.5 Historical Variation in the Histopathology of Lung Tumors.

The distribution of histopathologic types of all lung cancer cases has changed throughout the past century as evidenced in Zheng et al's (1994) and Thun et al's (1997) studies of Connecticut incidence data summarized in Figure 34. Use of these data to calculate risk parameters for the tracheal bronchial and peripheral bronchiolar regions is impossible without reasonable knowledge of the percentage of the anatomical location of these several

forms of lung tumors. Since such data were unavailable, estimates were bracketed for the initiation parameter by considering the cases wherein 25%, 50% or 75% of all lung cancer cases arising in the tracheal bronchial epithelium. (Table 5, example 1920s birth decades, European American males).

It has been argued that the historical decrease in the ratio of squamous cell to adenocarcinomas is due to deeper inhalation of smoke from filtered cigarettes first introduced in the 1950s. (Thun et al, 1997) On the other hand, the calculated parameter κ_h for smokers was found to be historically invariant (Table 5).

κ_h = expected number of new initiated precancerous lesions per year

$$= \kappa_{h,squamous} + \kappa_{h,adeno} + \kappa_{h,other}$$

$$= 2\tau^2 r_{i,squamous} r_{j,squamous} N_{squamous} (\alpha - \beta)_{squamous}/\alpha_{squamous} +$$
$$2\tau^2 r_{i,adeno} r_{j,adeno} N_{adeno}(\alpha - \beta)_{adeno}/\alpha_{adeno} +$$
$$2\tau^2 r_{i,other} r_{j,other} N_{other}(\alpha - \beta)_{other}/\alpha_{other}$$

'other' refers to lung tumor types other than squamous cell and adenocarcinomas. This presents a paradox: how can the total number of new initiated precancerous lesions per year remain constant, while the age-specific risk for squamous cell carcinomas decreases relative to the risk for adenocarcinomas?

Analysis of the Connecticut data is not possible with the model, since the data for ages < 40 is needed to estimate the growth rate of precancerous lesions, but some logic can be used to generate some hypotheses. The number of cells at risk for each tumor type is logically a constant, so potential changes in either mutation or precancerous lesion growth rates must be considered.

Scenario 1: Decrease in the mutation or precancerous cell growth rates of cells which can potentially give rise to squamous cell carcinomas

A means for κ_h to remain historically invariant, while the ratio of squamous cell carcinomas to adenocarcinomas decreases, is for a potential decrease in either the mutation rate or the precancerous lesion growth rate for cells at risk for squamous lung cancers to be offset by an increase in the mutation rate or the precancerous lesion growth rate for cells at risk for adenocarcinomas. Thun et al (1997) however demonstrates no discernible change in $\kappa_{h,squamous}$, as the historical changes in the slope of the incidence curves for squamous cell carcinomas, F_h $\kappa_{h,squamous}$, are proportional to the calculated changes in the fraction at risk for all lung cancers, F_h, (Table 5). Mutation rates in cells which can potentially give rise to squamous cell carcinomas have therefore remained historically constant. Furthermore, the estimate for Δ_h among male cohorts in Thun et al (1997) does not change significantly. Since this parameter is dependent on $(\alpha - \beta)$, the growth rate of the precancerous lesions, this rate has also not changed historically.

The differences in the incidence rate of squamous cell carcinomas between cohorts is therefore solely dependent on the respective differences in smoking prevalence, such that no discernible decrease is found in the rate of initiation of squamous cell carcinomas. This scenario therefore cannot explain why the ratio of squamous cell carcinomas to adenocarcinomas has decreased historically.

Scenario 2: The average age of onset of adenocarcinomas with respect to squamous cell carcinomas has decreased historically

The age-specific ratio of squamous cell carcinomas to adenocarcinomas can decrease if the age of onset for adenocarcinomas decreases with respect to the age of onset for squamous cell carcinomas.

Figure 69a illustrates this principle for two theoretical incidence curves, A and B, with a difference in the age of onset of 6 years. Figure 69b shows their respective ratio. When curve B is shifted left along the x-axis (equivalent to decreasing the average age of onset) this causes the ratio of the incidence rate of A to the incidence rate of B to decrease.

Thun et al (1997) shows that Δ_h for squamous cell carcinomas has not varied historically for men born in Connecticut, at about 50 years of age. On the other hand Δ_h for adenocarcinomas has decreased from about 55 years for the male cohort born in the 1890s to 45 years for the male cohort born in the 1930s. Plotting the ratios of the male incidence rates for squamous cell carcinomas to adenocarcinomas, the results resemble the expectation for two diseases with a decreasing difference in the age of onset. (Figure 69) A change in the age of onset of adenocarcinomas then likely explains the observed decrease in the decreasing ratio of incidence rates of squamous cell carcinomas to adenocarcinomas.

Possible hypotheses to explain a decrease in the age of onset of adenocarcinomas

If smoking prevalence data for men born in the US were representative of smoking prevalence in Connecticut, the observed change in Δ_h for adenocarcinomas is then not due to differences in the age at which individuals began smoking. Men born in the 1890s began smoking at an average age of 19 while men born in the 1930s began smoking at an average age of 17. (Harris, 1983) The change in Δ_h for adenocarcinomas (Figure 34) is however larger.

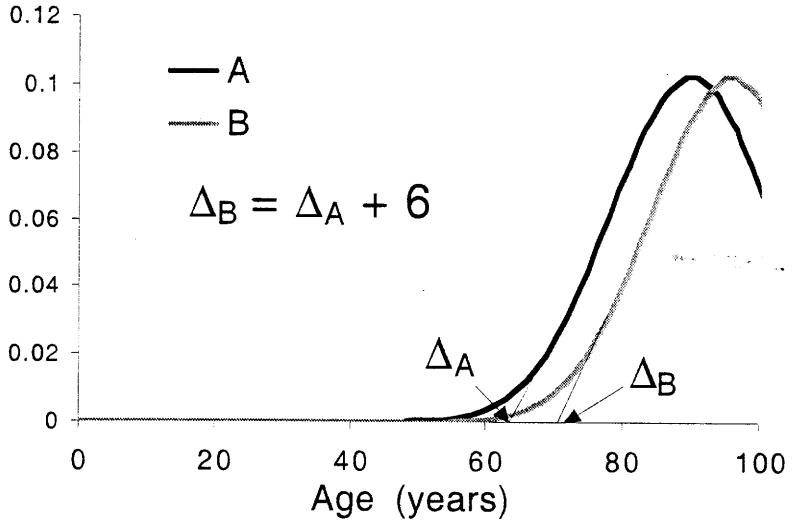
Other possibilities are that the age of onset was decreased by either (a) an increase in the promotion mutation rate of cells in the precursor lesion of adenocarcinomas, OR (b) an increase in the growth rate of the precursor lesions of adenocarcinomas.

Fig. 69: Effect of a change of onset of one form or death relative to another, on the ratio of their rates.

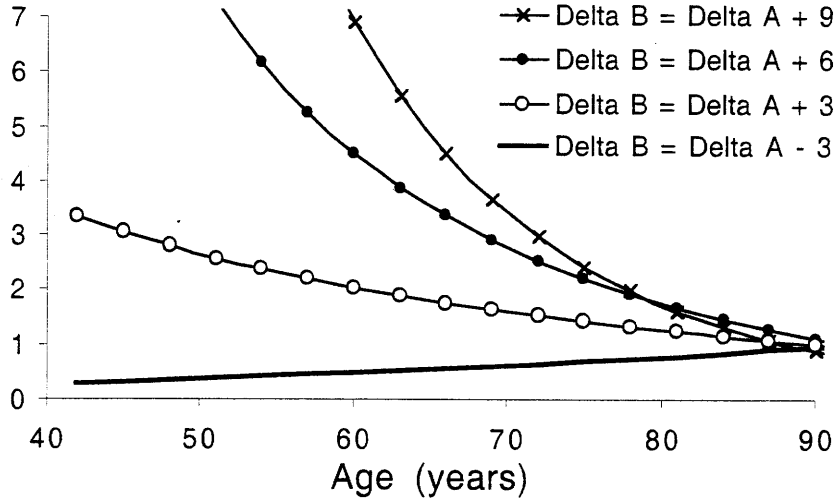
a) Theoretical incidence curves A and B with a difference of age of onset, $\Delta_B - \Delta_A$, of +6;

b) Ratio of incidence rates A and B, as a function of the difference in the age of onset $\Delta_B - \Delta_A$.

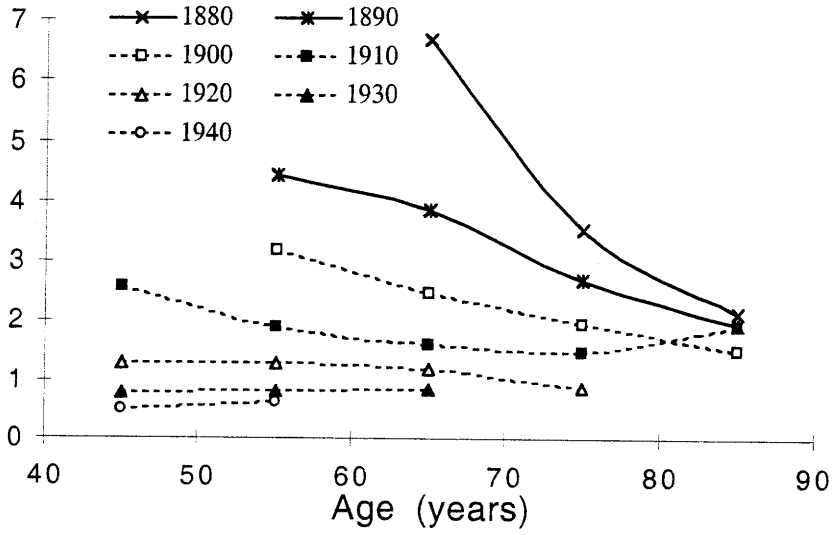
c) Ratio of incidence rates for squamous cell carcinomas to adenocarcinomas as a function of birthyear cohort. (Thun et al, 1997)


0.12
0.1
0.08
0.06
0.04
0.02
0
A
B
$\Delta_B = \Delta_A + 6$
Δ_A
Δ_B
0
20
40
60
80
100
Age (years)

a)


7
6
5
4
3
2
1
0
Delta B = Delta A + 9
Delta B = Delta A + 6
Delta B = Delta A + 3
Delta B = Delta A - 3
40
50
60
70
80
90
Age (years)

b)


7
6
5
4
3
2
1
0
1880
1890
1900
1910
1920
1930
1940
40
50
60
70
80
90
Age (years)

c)

For those cohorts for which few smokers consumed unfiltered cigarettes, exposure in the lower sections of the lung would be rare. If promotion mutation rates were affected by cigarette exposure, then incidence rates for adenocarcinomas would concurrently be low in these individuals, because they primarily arise in lower regions of the lung than do squamous cell carcinomas. The age of onset for adenocarcinomas would thereby be expectedly later in life. For more recent cohorts which began to use filtered cigarettes, these lower sections of the lung would receive higher exposures, theoretically driving up the promotion mutation rates. This would result in a decrease in the age of onset, which concurrently drives up the age-specific incidence of adenocarcinomas.

Wistuba et al (1997) have shown that about half of dysplastic colonies in smokers show an elevated LOH rate, whereas no discernible LOH was seen in respective dysplastic colonies in nonsmokers. If promotion of lung precancerous lesions were to consist of a series of LOH events, then cigarette smoking would expectedly decrease the age of onset of adenocarcinomas as it is inhaled more deeply with the advent of filtered cigarettes. Contrarily, the age of onset of squamous cell carcinomas would not be affected as these primarily occur in upper sections that were hypothetically already being exposed by cigarette smoke from unfiltered cigarettes. Thus, an increase in promotion mutation rates of adenocarcinomas could explain why the ratio of squamous cell carcinomas to adenocarcinomas has decreased, while the parameter κ_h has remained constant, since κ_h is independent of the rate of promotion mutation.

Alternately, the age of onset of adenocarcinomas can be affected by a change in the growth rate of precancerous lesions. Based on the results for individuals in the US (Table 5), the primary effect of cigarette smoke was on the rate at which precancerous lesions grow. These lesions were calculated to grow almost twice as fast in a smoker than in a nonsmoker.

This effect may be more limited in the lungs of smokers of unfiltered cigarettes, such that the precursor lesions of adenocarcinomas in these smokers grow more slowly due to limited exposure. Those pre-adenocarcinoma lesions with high exposure (smokers of filtered cigarettes) would grow at a faster rate than lesions with lower exposure (smokers of unfiltered cigarettes).

Such a potential scenario would result in an earlier age of onset of adenocarcinomas among smokers of filtered cigarettes compared to smokers of unfiltered cigarettes, which concurrently would drive up the age-specific incidence of adenocarcinomas as filtered cigarettes were introduced. Again, the age of onset of squamous cell carcinomas would not be affected since the upper regions where these tumors primarily arise would have been exposed by unfiltered cigarettes and filtered cigarettes alike. This would therefore explain the observed historical decrease in the ratio of squamous cell carcinomas to adenocarcinomas.

In this case, we must however note that the parameter κ_h is dependent on the rate at which precancerous lesions grow. In order to maintain κ_h constant while $(\alpha - \beta)_{adeno}$ increases historically, the division rate of cells in the precursor lesion of the adenocarcinoma, α_{adeno}, must increase by the same proportion as the death rate of cells in the precursor lesion of the adenocarcinoma, β_{adeno}. Increasing both division and death rates by the same proportion keeps the term $(\alpha - \beta)_{adeno}/\alpha_{adeno}$ and thereby the number of expected new precancerous lesions per year, κ_h, constant even with a change in the growth rate of the precancerous lesions, $(\alpha - \beta)_{adeno}$. In the analysis summarized in Table 5, calculations had been made assuming that α was the same for smokers and nonsmokers alike. To confirm this, it would be of interest for research to be conducted to determine the difference, if any, in the division rate of cells in precancerous lesions of smokers and nonsmokers.

4.2.6 Comparison of the Fraction at Risk for Lung Cancer and Smoking Prevalence

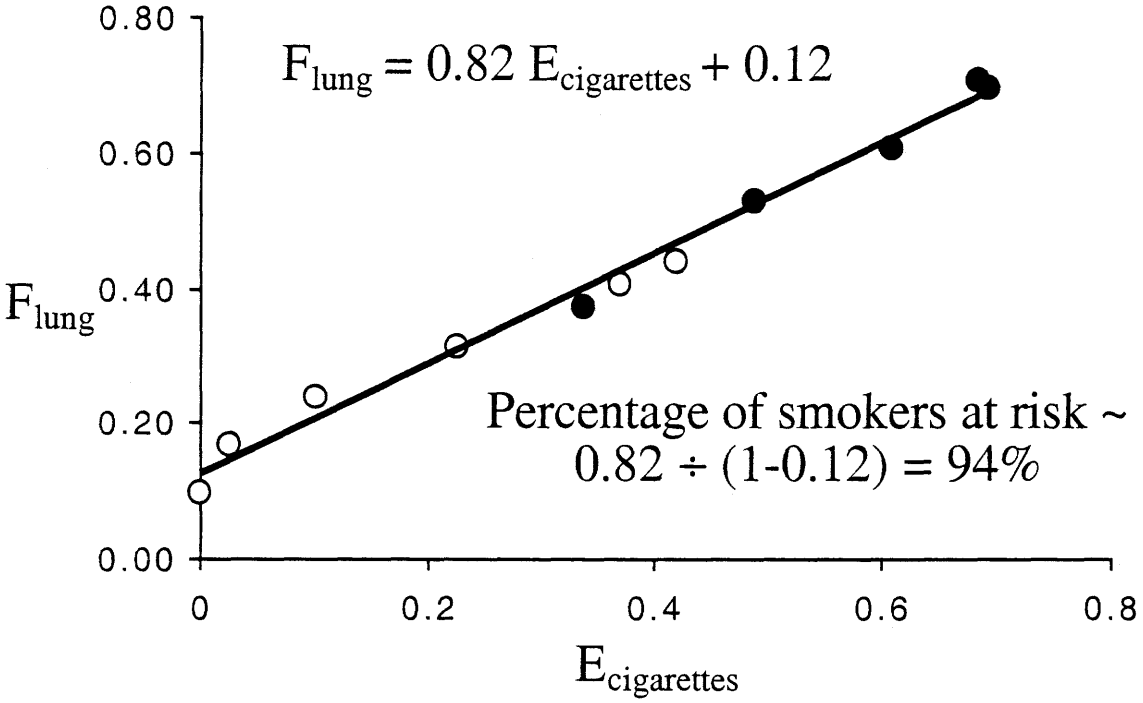
Plotting the estimates of the lifetime fraction at risk for each birth decade, $\mathbf{F_h}$, versus $\mathbf{E_{h,cigarettes}}$ (Figure 70), reveals a simple linear relationship. Strikingly, the data include all birth decade cohorts of European males and females as well as the estimate of lifetime lung cancer death risk for nonsmokers. Solving the implied algebraic relationship, the calculated fraction of smokers at lifetime risk of lung cancer is 0.94. However, the use of the maximum fraction of smokers at any age for each birth decade cohort to initially define $\mathbf{E_{h,cigarettes}}$ must to some degree include individuals of less than lifetime smoking habit and thus a lower lifetime risk. Thus $\mathbf{E_{h,cigarettes}}$ must to this degree represent an overestimate of lifetime smoking prevalence. Underreporting of cancer deaths in general and lung cancer in particular would, of course, lead to an underestimate of $\mathbf{F_h}$. and thereby an underestimate of the fraction of lifetime cigarette smokers at lifetime risk of lung cancer death. It is clear, therefore, that at least 94% of smokers are at lifetime lung cancer risk and a reasonable interpretation would include, and perhaps favor, an estimate of 100%. Lower estimates are not in accord with these data. (Belogubova et al, 2000).

4.2.7 Comparison of Age-Specific Lung Cancer Rates between Genders.

The idea that cigarette smoking women are at greater age-specific risk of lung cancer than cigarette smoking males has been suggested both by a number of studies in which cohorts consisting of female smokers with lung cancer were interrogated with regard to past smoking levels and studies of differential production of DNA adducts in the lung. (Risch et al, 1993; Ryberg et al, 1994; Kure et al, 1996; Tang et al, 1998)

Fig. 70: Calculated fraction at risk for lung cancer vs. smoking prevalence

Fraction of the population at risk for lung cancer as a function of maximum smoking prevalence for birth decade cohorts of European American males and females. (open circles - females born in the 1880s to 1920s; closed circles - males born in the 1880s to 1920s)


$F_{lung} = 0.82\ E_{cigarettes} + 0.12$
Percentage of smokers at risk ~
0.82 ÷ (1-0.12) = 94%
F_{lung}
$E_{cigarettes}$
0.80
0.60
0.40
0.20
0.00
0
0.2
0.4
0.6
0.8

Such cohorts are perforce smaller than the entire female population, but this aspect of smoking and cancer risk can be explored using a large cohort of males and females where the cohorts had identical self-reported maximum smoking prevalence. The fraction of adults using cigarettes was ~35% among males born 1881-1890 and among females born 1911-1920. (Harris, 1983). It is obvious by inspection that these populations showed near–identical age-specific lung cancer mortality rates. (Figure 71) These values include all men and women smokers in these birthyear cohorts, so that these data are perforce comparing the effect of average smoking habits in females to average smoking habits in males. They do not permit analyses of dose dependent effects of smoking levels or of other characteristics such as inhalation habits which may differ between males and females of these birth decade cohorts.

4.2.8 The Age-Specific Appearance of Dysplastic Lesions

The model can also be used to predict the expected age specific appearance of surviving initiated colonies. Figure 72 shows the predicted number of surviving preneoplastic lesions in the upper bronchial tracts of smokers and nonsmokers as a function of age. These expectations may be directly compared to the painstaking efforts reported by Wistuba et al (1997) in which dysplastic lesions and carcinomas *in situ* were enumerated in dissected upper airways of smokers and nonsnmokers of known age at death.

Wistuba et al (1997) reported zero moderate to severe dysplastic colonies or carcinomas *in situ* in twenty-one nonsmokers of average age of 29. For nonsmokers of this age the model predicts that fewer than 1 in 25 would carry a surviving preneoplastic lesion a prediction consistent with observation. (Figure 72) Wistuba et al (1997) also reported observations of 27 potentially preneoplastic colonies among 18 smokers born in the 1930s.

Fig. 71: Effect of gender on lung cancer mortality rates

Given equivalent maximum cigarette use

Males (1880s) and females (1910s).

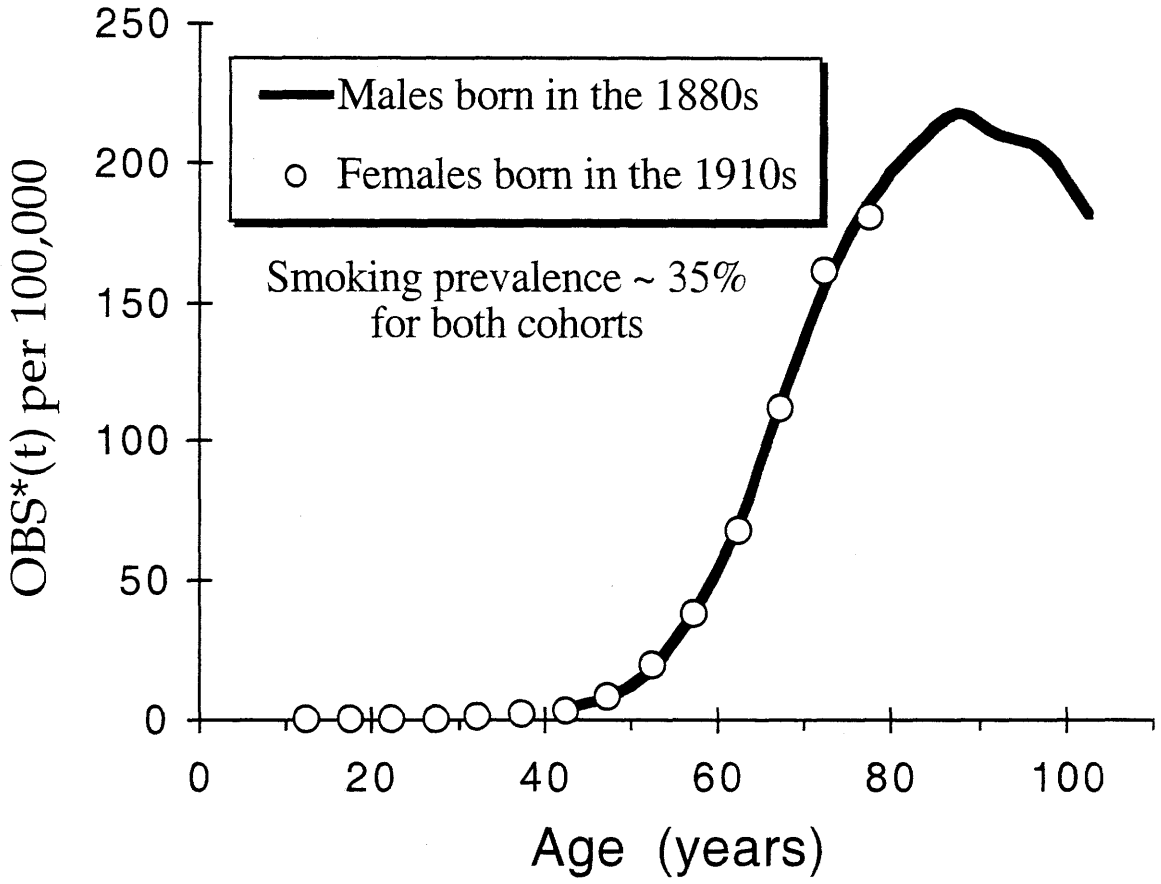
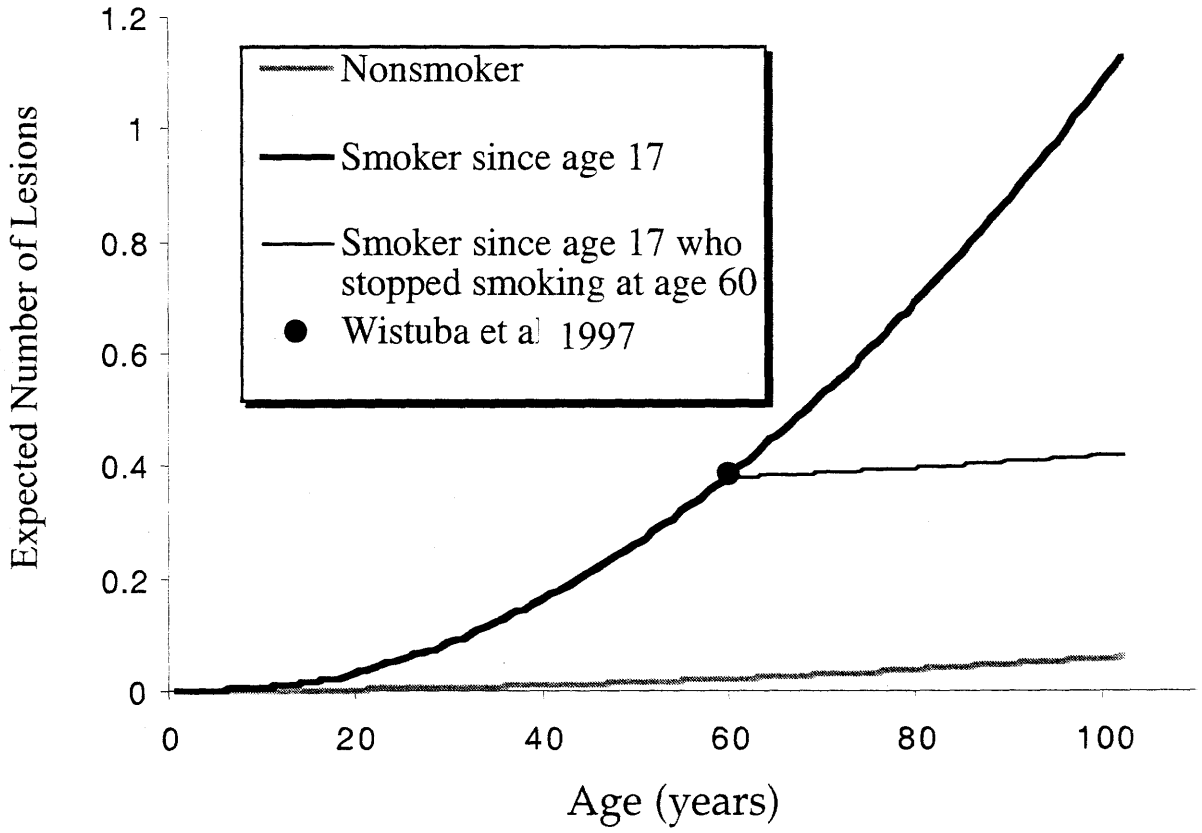

250
200
150
100
50
0
OBS*(t) per 100,000
Males born in the 1880s
Females born in the 1910s
Smoking prevalence ~ 35%
for both cohorts
0
20
40
60
80
100
Age (years)

Fig. 72: Expected number of preneoplastic lesions as a function of age

For these calculations it was assumed that 50% of preneoplastic colonies arise in the upper bronchial tree, that 10% of nonsmokers and 100% of smokers are at risk of developing preneoplastic colonies and that the growth rate in smokers is 0.32 and in nonsmokers 0.17, with identical initiation rates in smokers and nonsmokers, as recorded in Table 5. Shown as a filled circle is the number of carcinomas *in situ* per smoker (7/18) reported by Wistuba et al (1997) in a group of 18 male smokers of average age 60 yrs. Wistuba et al (1997) also reported a total number of moderate to severe dysplastic lesions of 1.1 per person (20/18) in this same group.


Expected Number of Lesions
1.2
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
Age (years)
Nonsmoker
Smoker since age 17
Smoker since age 17 who stopped smoking at age 60
Wistuba et al 1997

averaging 60 years of age. Of these, 7 were carcinomas *in situ* and 20 were moderately to severely dysplastic lesions. The predicted number of preneoplastic lesions in male smokers at age 60 is obviously consistent with the reported number of carcinomas *in situ* and significantly less than the number of other dysplastic lesions.

4.2.9 Prediction of Lung Cancer Mortality Rates Among Former Smokers

Peto et al (2000) published a large retrospective study of lung cancer in former smokers. They presented their data as the cumulative age-specific risk for cohorts of British males who smoked from mid teens or later to age 30, 40, 50, and 60. The cumulative age-specific risk is calculated as (1 – exp(-integral of $\mathbf{P_{OBS}(h,t)}$)) from birth to the age of interest. This provides a secondary test of the interpretation that the only two parameters affected by cigarette smoking were the fraction at lifetime risk, $\mathbf{F_h}$, and the growth rate of preneoplastic lesions, (α-β).

The approach is to calculate the size and distribution of preneoplastic lesions in smokers starting at the age of smoking initiation and ceasing to smoke at exactly ages 30, 40, 50 and 60 to accord as closely as possible with the conditions defined by Peto et al (2000). In physiological terms, existing preneoplastic colonies were assumed to continue to grow with a net growth rate of 0.17 doublings per year after smoking cessation, and that 10% of former smokers remained at lifetime risk of developing new lesions as was calculated for nonsmokers. (Table 5) This would intuitively have the effect of delaying the appearance of a tumor from a pre-existing preneoplastic lesion and would drop the number of expected subsequent surviving preneoplastic lesions to that expected in the average nonsmoker.

These predictions are presented along with the observations of Peto et al (2000) in Figure 73. The distributions of ages of smoking adoption and cessation among the several cohorts were not precisely known. Incidence among lifetime smokers best fit if the average age of smoking adoption was 22.5 years, which was applied to all other cohorts. The age smoking cessation had been defined as "around 60,50,40 and 30 years of age" leaving some uncertainty as to the actual length of smoking experience especially for those who ceased smoking at "around 30," but it was assumed that cessation occurred at exactly those ages. Since the epithelium of former smokers undergoes significant histologic changes back to that of normal epithelium of nonsmokers by five years after cessation, a relaxation time of two years to return to nonsmoker parameters is assumed. (Auerbach et al, 1962).

Figure 73 shows that the hypothesis that cigarette smoking affects growth rates of precancerous lesions is wholly consistent with Peto et al's (2000) data for cumulative risk of lung cancer among former smokers.

This is consistent with the analysis of the U.S. lung cancer mortality data which had suggested that growth rates of precancerous lesions in smokers were greater than in nonsmokers. Not having known these previous results, one may be tempted to test the alternate hypothesis, that cigarette smoke has a direct mutagenic effect on mutation rates of lung epithelial cells. Figure 73 predicts the effect that a decrease in mutation rates (both initiation and promotion) would have had on incidence among former smokers. The results conclusively agree better with the hypothesis that growth rates and not mutation rates are altered by cigarette smoke, but the possibility that there might be a small effect on mutation rates within smaller subpopulations of smokers (i.e. heavy smokers) could not be dismissed with the analysis alone of Peto et al's (2000) lung cancer incidence data among former smokers.

4.3 Caveat

Calculations based on the actual public health records and the assumptions outlined herein have one very important characteristic when it comes to thinking about mutation rates, cell kinetics and cancer. They are derived from observations of persons who died of cancer. They provide absolutely no information about mutation rates and cell kinetics in persons who are not at risk.

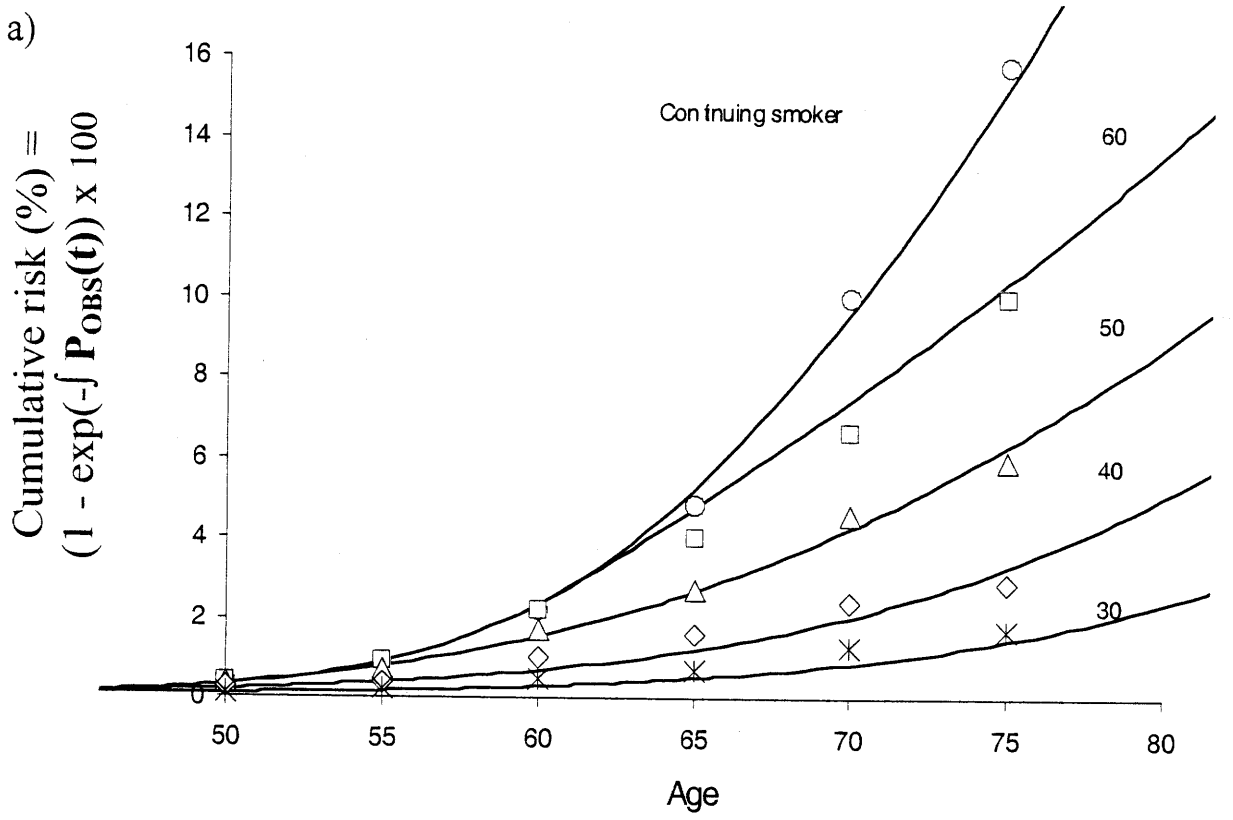
Fig. 73: Age-specific cumulative risk of developing lung cancer among former smokers

Markers represent data from Peto et al (2000) and lines represent the cumulative risks predicted by model. Numbers on the right hand side represent approximate age at which smoking cessation occurred.

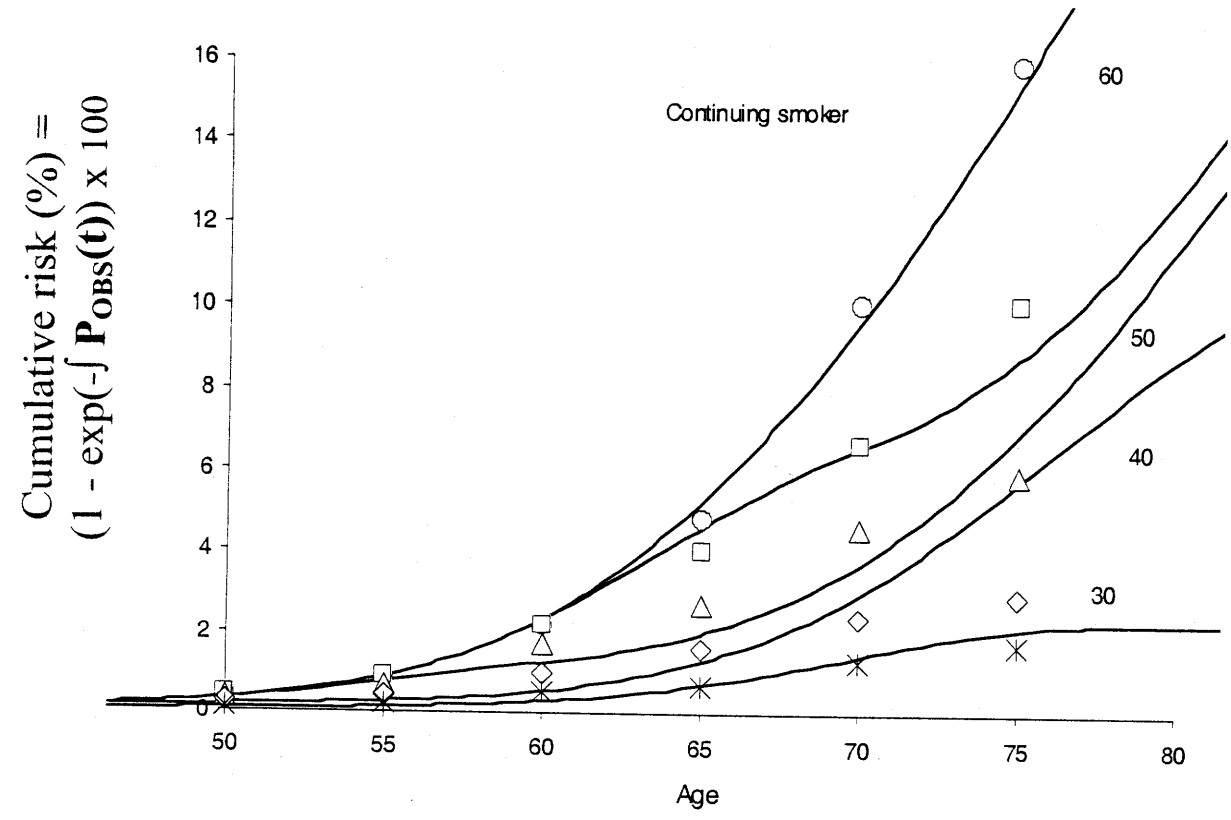
a) Predictions assume that preneoplastic lesions of current smokers grow at a doubling rate of 0.32, preexisting precancerous lesions in former smokers continue to grow at a doubling rate of 0.17 two years after smoking cessation, and only 10% of former smokers develop new precancerous lesions after smoking cessation, as calculated for nonsmokers (Table 5).

b) Predictions assume that preneoplastic lesions of smokers and nonsmokers grow at the same doubling rate of 0.32. Two years after cessation, initiation and promotion mutation rates were decreased 6-fold (best fit).

a)


Cumulative risk (%) =
(1 - exp(-∫ P_OBS(t)) x 100
16
14
12
10
8
6
4
2
0
Continuing smoker
60
50
40
30
50
55
60
65
70
75
80
Age

b)


Cumulative risk (%) =
(1 - exp(-∫ P_OBS(t)) x 100
16
14
12
10
8
6
4
2
0
Continuing smoker
60
50
40
30
50
55
60
65
70
75
80
Age

5. DISCUSSION

5.1 General Conclusions

Use of the data for survival probabilities as a function of age and history has extended and confirmed the earlier interpretation based on mortality data alone that there is a true maximum in the age-specific colon cancer mortality rate for males and females in both European- and African-American subpopulations. This interpretation is essential for the development of a means to calculate the fraction at risk of cancer for each birth year cohort.

Using an extended Knudson-Moolgavkar model for initiation and promotion, in which 'n' rare events are required for initiation and 'm' for promotion, and the approximation of Equation 14 (Section 3.5.2.2) to account for competing forms of death sharing environmental and/or genetic risk factors with cancer, permits calculation of birth year cohort-specific values for the fraction at primary risk, the product of initiation mutation rates, the product of promotion mutation rates, and the average growth rate of the precancerous intermediate colony.

Since these parameters have been calculated for each birth year cohort, their historical changes may be observed. These in turn may be considered in terms of historical changes in human habits and their environment. These parameters may be compared between the two large demographic cohorts for which data are available. Similarly, the parameters for males and females may be compared.

In particular, the robustness of the estimates for each parameter calculated relative to uncertainties in the accuracy of the recorded data has been determined. The effects of +/- 10% errors in the estimates of the slopes of the ascending age-specific mortality data or the intercepts of these data extrapolated to an age intercept were examined on estimates of the calculated parameters. These uncertainties were found to result in errors of less than +/- 25% for estimates

of initiation mutation rates and growth rates of preneoplastic lesions. However, errors in the estimates of the rates of promotion events were far less robust leading us to place little reliance on what might otherwise have been interpreted in historical shifts in this parameter. In short, estimates derived from the data set by this mode of analysis are no better but no worse than the public record and the accuracy of the three-stage hypothesis of carcinogenesis that underlies our mathematical model.

In one important area, the use of estimates of mortality rates at ages above 80 years, the degree of accuracy trespasses on the widely held belief that such data are grossly inaccurate and unusable for quantitative analyses. Recent records are bolstered by the prevailing Medicaid/Medicare system of reporting the care and causes of death in the very aged and are in general agreement with earlier records of death rates at ages > 80 yrs. Furthermore, mortality rates among the elderly in Malmo, Sweden, where autopsies were once required after every death, also decreased among the elderly. (Prof. Thilly, personal communication)

Even so only the accuracy of but one parameter, $\mathbf{f_h}$, would be seriously affected by the supposed underreporting of death rates among the extremely aged for any specific cancer type. Such underestimates would lead to overestimation of $\mathbf{f_h}$, the fraction of the lifetime population at risk that is expected to die of the particular cancer given only multiple competing forms of mortality. (Figure 52)

5.2 Colon Cancer

5.2.1 Historical Changes in the Fraction at Primary Lifetime Risk, $\mathbf{F_h}$.

Implicit in the assumption that an individual is or is not at lifetime risk of any particular form of cancer is the existence of inherited and/or environmental factors that create a condition essential for the incidence of cancer during a maximum human life span. An inherited factor necessary for initiation or promotion that is not carried by all members of the population is an essential factor if persons without this inherited factor would not experience the specific cancer type in a maximum human life span. An environmental factor could take the place of an inherited factor in terms of physiological effect or it could interact in an essential manner with an inherited essential risk factor.

The fraction at primary risk of colon cancer has remained essentially constant for the birth year cohorts of the 1860s to the 1940s (Figure 61). This is true for males and females of both European or African heritage. This fraction is about 0.4. It is possible that this fraction at risk was increasing from 0.3 for the birth cohort for the 1840s to 0.4 for the 1870s.

The constancy of this fraction during a period of marked changes in American life in nutrition, smoking habits, level of exercise, industrialization and urbanization is striking. These data suggest that none of these known environmental changes had any effect on the fraction at risk. It might be imagined that there have been offsetting environmental changes but such arguments, absent data, violate the "law of parsimony". It should be noted that this result does not indicate that there are no environmental factors affecting age-specific colon cancer rates. Subpopulations with conditions varying significantly from the population average might have higher or lower rates depending on their circumstances.

A large fraction at risk of 0.4 may seem surprising, given that less than 5% of all deaths result from colon cancer. This result, however, emphasizes the importance and necessity of accounting for all other connected forms of death in calculating the primary risk fraction for any mortal disease, as well as taking into consideration that an individual dying of an independent form of death (i.e. accident) could have potentially died of colon cancer had they not prematurely died.

The estimate of 0.4 represents a minimum value for the fraction at primary genetic risk. If all persons were at environmental risk (i.e. no environmental risk factor necessary), then the fraction at primary genetic risk would be 0.4.

Case I: Genetic risk is conferred by a dominant mutation <u>non-deleterious</u> for reproductive fitness

In this case, homozygous recessives (wild type) would not be at genetic risk, but heterozygotes and homozygous dominants would be at equal risk. If risk were monogenic and the dominant and recessive alleles have reached Hardy-Weinberg equilibrium in the population, then the sum of heterozygous and homozygous dominant fractions would be $0.4 = 2pq + q^2$. "p" is the allele frequency of recessive and "q" of dominant alleles such that $p + q = 1$. Solving this quadratic equation, $q = 0.23$.

For the case of a dominant monogenic primary genetic risk factor, the sum of inherited alleles coding for risk would thus be 0.23. For multigenic risk the average value of q would be equal to (0.23 / number of genes), and for polygenic risk the average value of q would be equal to $0.23^{1/\text{number of genes}}$ (i.e. 0.5 for 2 genes).

These estimates are in the realm of possibility if there were no physiological effect on reproductive fitness for homozygous or heterozygous states. The physiological effect would be limited to a risk of death by colon cancer at advanced age. Since the average rate of gene mutations leading to gene loss is about $3x10^{-5}$ per human generation and there have been about 10^4 human generations, the accumulated mutant fraction of about 0.3 would be expected for the sum of a set of neutral alleles for a single gene.

A hypothesis that primary genetic risk for colon cancer is defined by any of a set of non-deleterious dominant mutations in one or several genes is thus not inconsistent with the calculated primary risk fraction of 0.4. The physiological effect of such a dominant mutation could affect initiation, promotion or progression, there being no way to differentiate among these possibilities with the existing data or understanding of carcinogenesis.

Case II: Genetic risk is conferred by homozygosity for a recessive mutation <u>non-deleterious</u> for reproductive fitness.

In the case where primary genetic risk for colon cancer requires inheritance of two recessive alleles of the same gene, neither of which affect reproductive fitness, the fraction of recessive homozygotes would be $q^2 = 0.4$, or $q = 0.63$ for a monogenic disorder.

Since these recessive alleles in homozygous or heterozygous form would have by definition no effect on reproductive fitness, they might have reached so high a fraction in present day populations if the mutation rate for a single gene were about twice the average for all gene inactivating mutations (See Case I) or if the risk were distributed over several different genes. As

in Case I, one could not logically deduce which stage of carcinogenesis might be affected by the recessive homozygous state.

Case III: Genetic risk is conferred by a recessive mutation deleterious for reproductive fitness.

A third possibility is that risk is conferred by a set of alleles in one or more genes in which homozygosity for such mutations is lethal in embryos or at least prevents reproduction. Assuming again that these alleles are in Hardy-Weinberg equilibrium and that mutations leading to gene loss, average about $3x10^{-5}$ per generation, the average expected sum of mutant allele fractions for heterozygotes in any one gene would be 0.013 in the population. The actual value for any gene would depend on gene size and the presence of particularly marked mutational hotspots. For these to sum to 0.4, a multigenic model is obviously required. Forty separate genes each at the Hardy-Weinberg equilibrium value of about 1% would be a hypothesis consistent with these calculations.

In colon cancer, it would appear that any required event involving loss of heterozygosity would occur during promotion since the events of initiation are accounted as loss of two wild type APC alleles, and events in progression would not be rate-limiting. An inherited condition of heterozygosity in one of these 40 genes would therefore be sufficient to account for primary risk since the need for only a single LOH/LOI event for promotion would presumably place these individuals at equal risk.

A model considering a polygenic combination of deleterious recessive alleles would have to consider a very large number of genes (>1000). This consideration leads to the conclusion that a combination of LOH events during promotion involving two or more genes carrying alleles deleterious for fitness is an unlikely scenario.

In summary, a primary genetic risk fraction of 0.4 or higher could be conferred by mutant alleles of one or a few genes if reproductive fitness were not affected. On the other hand, 40 or so genes would seem to be required to create so high a primary genetic risk fraction if homozygosity for the mutant alleles did prevent reproduction. In the former case, the original alleles occurring as hotspot mutations would have arisen and been fixed multiple times throughout human history. In the latter case, selection against ancient deleterious mutations would leave only relatively recent mutations. In large present day populations, such as those of Asia, Africa and Europe, these would be expected to be distributed over very large numbers of families.

5.2.2 Historical Changes in the factor Accounting for Connected Risks, $\mathbf{f_h}$.

As may be seen in Table 1 (Section 4.1.1), $\mathbf{f_h}$ increases markedly in historical time for three of the four cohorts and has a maximum value of 0.24 in the most recent European-American male cohort for which $\mathbf{f_h}$ may be calculated. Values for males are generally higher than for females.

This factor accounts for both underdiagnoses, underreporting, and deaths of persons at risk of colon cancer by other diseases which share the genetic and/or environmental risk factor(s). Underdiagnoses and underreporting should have decreased from 1930 to 1992. The increasing value of $\mathbf{f_h}$ is probably in part accounted for by this trend. On the other hand, the low value of $\mathbf{f_h}$ derived from the populations born in this century suggests that the genetic and/ or environmental risk factors for colon cancer are responsible for a significant fraction of other

deaths. Given that colon cancer accounts for somewhat less than 5% of all deaths and that the value of $\mathbf{f_h}$ is about 0.2, one must consider that risks for colon cancer are associated with as much as (5% / 0.2) = 25% of all deaths. So large a fraction could comprise all cancer deaths; alternately, the genetic or environmental risk factors for colon cancer could contribute to a large fraction of vascular disease.

As noted in Section 3.5.2.2, $\mathbf{f_h}$ is an approximation forced by ignorance of any forms of death sharing risks with colon cancer. It is the shakiest part of this modeling effort and represents an area in which more theoretical work is needed.

5.2.3 Historical Changes in Initiation Mutation Rates, $\mathbf{r_i}$ $\mathbf{r_j}$.

Estimates of the rate of the first initiation mutation for the condition n=2 varied from 4 to 8×10^{-8} over all four gender and ethnic cohorts for the birth year cohorts from the 1840s to 1940s. Based on Grist et al (1992) the ratio of the loss of an active gene by primary mutation was approximately one third the rate of allelic loss by LOH. Thus, $\mathbf{r_j} = 3\ \mathbf{r_i}$ has been used to calculate $\mathbf{r_i}$ after the product $\mathbf{r_i r_j}$ was calculated. These values are remarkably similar to observed rates of spontaneous mutations for gene inactivation in human cell cultures of about 10^{-7} per cell division. They are almost identical to an estimate of about 0.7×10^{-7} per stem cell division derived from the age dependent hprt mutant fractions in human peripheral T cells assuming three stem cell divisions per year. (Figure 8) (Bigbee et al, 1998; Branda et al, 1993; Davies et al, 1992; Finette et al, 1994; Henderson et al, 1986; Hirai et al, 1935; Hou et al, 1995; Hutner et al,

1995; Liu et al, 1997; McGuniss et al, 1990; Tates et al, 1991). These values are consistent with a model of loss of the first APC allele in colonic stem cells at a rate of about 7×10^{-8} per stem cell division and a rate of LOH for the second allele at a rate of about 2.1×10^{-7} per stem or transition cell division. So close are these calculated values to observed human *in vivo* mutation and LOH rates that n = 2 is established for colon cancer initiation until contradictory evidence is discovered.

The mutation rate for European-American Females is essentially invariant at 7×10^{-8} with historical time, but the data suggest a significant increase in mutation rates in both male cohorts from a steady value of 4×10^{-8} from the 1840s through 1880s to over 6×10^{-8} from the 1880s through the 1940s. African-American Females appear to show a steady increase from 4×10^{-8} in the 1840s to the 1900s when it reaches the rate of 7×10^{-8} seen in European-American Females.

Considering each birth year cohort as an independent trial, one could agree that the differences are statistically significant. However the accuracy of primary data for OBS(h,t), S(h,t) and R(h,t) cannot be ascertained, nor can the accuracy of the approximation represented by $\mathbf{f_h}$. The apparent differences might arise from unknown biases.

Another uncertainty is that the studies of Fuller et al (1990) and Jass et al (1992) provide estimates for the rates of colon stem cell LOH of about 7×10^{-6} LOH events per colonic stem cell division, 30 times higher than LOH rates derived from observations in T cells. If these LOH rates for colonic stem cells are accurate and apply to transition cell divisions, then the estimate of

the loss of the first APC allele by point mutation or deletion must be reduced to about 2×10^{-9} per stem cell division. There are no data available to exclude this possibility.

5.2.4 Historical changes in Promotion Mutation Rates, $\mathbf{r_A}$.

The number of genetic changes required for promotion in colon cancer is unknown. Values for the geometric mean of the mutation rates for m=1, 2, 3 .... as an estimate of $\mathbf{r_A}$ must be considered.

These genetic changes could be gene "activation" missense mutations, gene inactivation events, LOH for an inherited heterozygous state, or loss of imprinting of a gene by other mechanisms (LOI). These processes could involve point mutations, recombination, chromosomal and or chromosomal segment loss. As noted above (Section 5.1.1), on the basis of population genetics, there could be one and only one promotional LOH event in the case of an inherited recessive allele deleterious for fitness in the inherited homozygous state.

For the case m=1, the estimated value of $\mathbf{r_A}$ is about 2×10^{-7} per cell division for females and 8×10^{-8} in males, a value which is approximated by LOH in human T cells *in vivo* or gene inactivation of a somewhat larger than average gene. It is however much lower than the colonic stem cell LOH rate of 7×10^{-6} derived from Fuller et al (1990) and Jass et al (1994). If any of multiple genes were involved, then activation of any of several proto-oncogenes might also be considered a numerically reasonable hypothesis.

It is clear that if m=1 no increase in promotional mutation rates above those seen in normal human T cells need be invoked to account for the age-specific colon cancer rates in

humans. Curiously, the historical estimate of this promotion mutation assuming m = 1 is remarkably constant for both European and African-American males at about 8×10^{-8}. For both European-American and African-American females it appears to have risen significantly from the mid-ninetieth century to a constant level of about 2.5×10^{-7} since the 1890s.

The differences between the genders and similarities between the ethnic groups may give some reason to place confidence in these results, leading to the question of what environmental changes may have affected all women beginning in the 1860s that was completed by the 1890s which might conceivably have affected promotional mutation rates. On the other hand, the differences, while apparent, may have arisen by the action of unknown biases in reporting or diagnosis which were in some way gender specific. Given the economic differences of the ethnic groups, however, one would have expected such biases to effect comparison between ethnic groups. Such differences do not appear at all.

5.2.5 Historical Changes in the Growth Rate of Precancerous Lesions, $(\alpha - \beta)$.

These values are extraordinarily constant at about 0.2 for males and 0.17 for females over the entire historical period analyzed. The gender specific differences appear to be real and constant over a century of birth year cohorts. There appear to be no differences between the two ethnic groups. It would appear that the many environmental changes during the century observed have had no effect on the net growth rate of colon adenomas.

It is worth noting that these net growth rates of 0.2 and 0.17 doublings per year are remarkably similar to the net growth rates of children which are about 0.16 (Figures 35 a,b). Net

colon carcinoma growth rates are about 20 doublings per year which may be compared to the net growth rate of the human fetus of some 54 doublings per year.

These similarities give a quantitative basis for the idea that the genetic steps of. carcinogenesis recreate the conditions of fetal and postnatal growth in reverse. In this scenario, the mutations permitting adenomatous growth take the cell back to the growth rates of children while the additional change(s) creating a carcinoma cell permit the more rapid growth rate of fetal life.

It is necessary to note that the observations of α, β and τ were made in adult colons. It would be interesting to know if τ changes in neonatal and childhood growth.

5.2.6 High LOH and LOI Levels in Human Colon Carcinomas.

The fraction of loci showing LOH or LOI is on average about 0.22. This high fraction would not be produced in adenomatous growth at the LOH rate observed in human T cells of about 2×10^{-7} mutations per cell division. The rate of LOH/LOI necessary to achieve such a fraction was calculated out to be about 4×10^{-4} per cell division. A comment on a common error in this matter of high LOH/LOI levels in tumors is in order. Some cancer researchers have used only the number of net doublings in adenomatous growth to account for an LOH/LOI fraction of 0.22. This would be the $\log_2$ of the cell number of an average adenoma at promotion, which is $\log_2 \frac{\alpha}{\alpha-\beta} 2^{(\alpha-\beta)\Delta_h}$, or about 17. However one requires the total number of linear divisions between the first adenoma and first carcinoma cell for this calculation. This number is about $\alpha\Delta_h = 567$.

This represents an estimated 2000 fold greater rate than observed in normal human lymphoid cells *in vivo* and *in vitro* and about 60 fold higher than estimated for normal colon stem cell LOH rates. So high a rate would however accommodate a value of m = 4 LOH/LOI events required for promotion (Table 3) in initiated cells.

Even if this high LOH and LOI rate occurred in colon adenomas, it would not necessitate the conclusion that the number of promotional events were 3 or 4 or that LOH or LOI were involved in promotion. Even if the LOH/LOI rates increased to 4×10^{-4} in adenoma cells, the necessary promotion event might still be a single point mutation occurring at a rate of 2×10^{-7} per cell division.

5.2.7 Conclusions

From public records from the 1930s to the present day, the three-stage carcinogenesis model predicted that the fraction of the population at primary risk for colon cancer risk was historically invariant at about 42% for the birth year cohorts from 1860 through 1930. This was true for each of the four demographical groups examined (European- and African-Americans of each gender). Additionally, the data indicate an historical increase in the initiation mutation rates for the male cohorts and the promotion mutation rates for the female cohorts. Interestingly, the calculated rates for initiation mutations are in accord with mutation rates derived from observations of mutations in peripheral blood cells drawn from persons of different ages. Adenoma growth rates differed significantly between genders but were essentially historically invariant.

A historically variant calculated fraction at risk would have insinuated that there had been a historical change in either the extent or types of environmental exposures that place an individual at risk for cancer. On the other hand, the observation of a historically constant calculated fraction at risk for colon cancer insinuates that genetic risk factors alone predispose an individual to colon cancer, but does not exclude the possibility that environmental risk factors still play a role in accelerating the time of occurrence of colon cancer within smaller subpopulations within the fraction at risk.

Brindha Muniappan's (unpublished) study of the induced *in vitro* mutational spectrum of the APC gene by replication with DNA polymerase β, has shown that 3 of the 6 so far characterized, induced mutational hotspots are concordant with the mutational hotspots found in the APC gene of cells taken from colon tumor samples. As the loss of function of the APC gene has been shown to be necessary, but not yet shown to be explicitly sufficient, in the initiation of at least 80% of all colon tumors (Powell, 1982), Brindha Muniappan's results are consistent with the hypothesis that primary risk factors for colon cancer are endogenous.

Furthermore, the three-stage carcinogenesis model predicted a potential two-fold increase in the calculated initiation mutation rates of colonic cells in males (Table 1) during the last century, insinuating that environmental risk factors, presumably changes in the diet of the average male, may play some role in accelerating the occurrence of colon cancer. However, these calculations were made assuming that the turnover rate of normal colonic tissue has been constant throughout this same period. Our calculations can in fact not distinguish between these two phenomena; only the product of the turnover rate τ and the mutation rate r_i could be estimated. As such, Kinzler and Vogelstein's (1996) hypothesis that dietary factors may play a

role in carcinogenesis by acting as irritants that accelerate tissue regeneration, is consistent with the results of the three-stage carcinogenesis model.

Likewise, the three-stage carcinogenesis model predicted a potential two-fold increase in the calculated promotion mutation rate of colonic cells in females (Table 2) during the last century. However, the robustness of the model (Section 4.1.6) demonstrates that the uncertainty in the calculated parameter of the promotion mutation rate, $\mathbf{r_A}$, alone could explain this calculated result. Possible unknown biases in the primary data sets of mortality, survival, and underreporting could explain the calculated changes in promotion mutation rates.

Lastly, the three-stage carcinogenesis model predicted a historical change in $\mathbf{f_h}$, the ratio of the number of deaths attributed to colon cancer to the number of all deaths sharing the risk factors of colon cancer. If either treatment for the connected diseases, which share the same risk factors as colon cancer, had improved more rapidly than treatment for colon cancer during the last century, or the number of underdiagnosed colon cancer deaths had decreased historically, the increased risk of dying and actually being correctly reported of dying of colon cancer within the elderly population would explain the calculated increase in $\mathbf{f_h}$.

Together, these calculations and observations suggest that the observed changes in colon cancer mortality rates are predominantly explained by the increase in the lifespan of the American population. New environmental exposures during the last century did not appear to place more individuals at risk. Instead, environmental risk factors may potentially have a secondary role accelerating the occurrence of cancer at initiation, by either increasing the rate of normal tissue turnover or by increasing the rate of mutation by means other than through a direct

mutagenic effect (i.e. impairment of the mechanisms by which mispaired bases in the DNA are repaired).

5.3 Lung Cancer

5.3.1 Historical Changes in the Fraction at Primary Lifetime Risk, $\mathbf{F_h}$.

The argument that there exist genetic and/or environmental factors that create a condition essential for the incidence of cancer during a maximum human life span has been anticipated by others who previously noted a maximum age-specific rate for several cancer sites. (Cook et al, 1969; Smith, 1996). Smith (1996), for instance, noted a decrease in mortality rates among the elderly for lung cancer and commented, "If most smokers die between ages 65-70, lung cancer mortality rates would expectedly decrease as the subpopulation at risk is killed off at a faster rate than the remaining non-smoking population."

The derived estimate of the fraction of lifetime nonsmokers at risk is ~10%. The fraction of nonsmoking persons at risk does not appear to have changed since the mid-1800s to the present day as may be concluded by inspection of Figure 66. This resembles the finding that $\mathbf{F_h}$ was unchanged for colon cancer in the American population since the birth years of the 1860s. Given that inherited risk has not significantly changed in the human population in so short a span of years, the many environmental changes between the mid 1800s and mid 1900s are concluded to have had little if any net effect on lifetime lung cancer risk in nonsmokers. The data however do not provide any indication as to the nature of the environmental and/or inherited essential lifetime lung cancer risk factors in nonsmokers.

In the case of lung cancer for cohorts of mixed populations of smokers and nonsmokers, there is an unambiguous linear relationship (Figure 70) between smoking prevalence and the

fraction of the total population calculated to be at lifetime risk, $\mathbf{F_h}$. Data for males and females of Americans of predominantly European descent fall on this same straight line. This quantitative relationship is evidence that cigarette use is an environmental factor essential for lifetime risk lung cancers in smokers. Were cigarette smoking to affect age-specific mortality rates solely by accelerating either initiation and/or promotion by inducing higher rates of required genetic changes, then lifetime risk fractions $\mathbf{F_h}$, would be the same for smokers and nonsmokers and would also be historically invariant, as was the case for colon cancer.

The linearity of the relationship permits the inference that cigarette use alone is sufficient to account for lung cancer primary risk in male and female smokers. The hypothesis that some other essential environmental factor for lung cancer in smokers, e.g. air pollution, rose simultaneously with cigarette use is not supported by the marked linearity of the relationship. Data from men and women born many decades apart would not yield the same value for $\mathbf{F_h}$ given the same prevalence of cigarette use if such were the case. (Figure 71)

The fraction at risk was calculated assuming zero survival for lung cancer, and that the maximum smoking prevalence throughout the lifetime of a cohort was representative of the smoking experience of that cohort. An underestimate in the true survival rate for lung cancer or an overestimate of the true average smoking experience have the net effect of causing an underestimate in the lifetime risk fraction, so a reasonable estimate of the fraction of lifetime smokers at lifetime risk of lung cancer is ~100%. This suggests that there are probably no lifetime smokers who would escape death by lung cancer if they did not die by some other cause i.e. there is no inherited lifetime resistance to cigarette induced lung cancer. Interestingly, the historical changes in the histopathological nature and anatomical location of lung tumors in

smokers, Figure 34, has had no effect on the value of $\mathbf{F_h}$ for smokers, meaning that a smoker would be at risk for all forms of lethal lung cancer.

Although some lifetime smokers do reach extreme old age without lung cancer, this is not inconsistent with the finding that all smokers are at lifetime risk for lung cancer. Initiation and promotion rates could be similar among all smokers but stochastically distributed such that some smokers die of lung cancer earlier than others. Surviving smokers have simply not lost yet in a game ruled by chance, or die by chance from another form of death before developing lung cancer.

5.3.2 Historical Changes in the Factor Accounting for Connected Risks, $\mathbf{f_h}$

Introduction of the parameter $\mathbf{f_h}$ was required by the seemingly obvious fact that persons at lifetime risk of lung cancer would frequently die of some other disease for which cigarette smoking was also a risk factor. Basically $\mathbf{f_h}$ is the fraction of persons at lifetime risk of lung cancer who would die of lung cancer if deaths were caused only by those factors essential for lung cancer, i.e. not including deaths by accident, infectious disease, etc. Its use in a mathematical model of age-specific cancer mortality is essential. The persons at risk of lung cancer would have higher age-specific death rates than those not at risk not only because of lung cancer deaths but also because of deaths caused by diseases having the same essential risk factors as lung cancer.

$\mathbf{f_h}$ for nonsmokers was 0.15 (Table 5). This indicates that the essential inherited and/or environmental factors required for lung cancer in nonsmokers are capable of killing by other forms of mortality in a manner quantitatively similar to smokers as noted below. Since

nonsmokers' deaths by lung cancer account for about 0.7% of deaths of a birth year cohort, the fraction of all deaths attributable to the essential factors for lung cancer in nonsmokers may be estimated to be 4.7%. This is hardly an insignificant fraction of a birth year cohort and further analyses to seek these essential risk factors in nonsmokers would seem to be important.

Doll and Peto (1976) and Kahn (1966) studied cohorts of smokers and nonsmokers and measured the effects of cigarette smoking on multiple forms of mortality. Ischaemic heart disease risk, a common form of death among nonsmokers, was in fact increased twice as much as lung cancer risk in smokers. Of the excess total deaths due to cigarette smoking, only 15-17% was attributable to lung cancer. (Doll and Peto, 1976; Kahn, 1966) The estimate that 15-17% of excess deaths among smokers are due to lung cancer would be a reasonable approximation to $\mathbf{f_h}$ in the sense employed.

Calculated values for $\mathbf{f_h}$ range from 0.11 to 0.13 for female smokers and from 0.14 to 0.28 for male smokers (Table 5). The average value of $\mathbf{f_h}$ for all female cohorts analyzed (1880s-1920s) is 0.13 and for males (1870s-1920s), 0.20. The average for both genders and all cohorts is 0.165. These numbers are thus in reasonable agreement with the epidemiological estimates based on British males born in the 1900s (Doll and Peto, 1976) and U.S. males born in the 1910s. (Kahn, 1966)

5.3.3 Historical Changes in Initiation Mutation Rates, $\mathbf{r_i}$ $\mathbf{r_j}$.

For the purposes of simplification and by analogy with colon cancer, calculation of initiation mutations rates was made for the case of n = 2 necessary initiation mutations as

recorded in Table 5. Recognizably, however, there is no experimental data suggesting n = 2 for lung cancer in smokers or nonsmokers.

Since the upper bronchial tree has but 2.4×10^8 cells and the lower bronchiolar region about 2.8×10^{10} cells, these anatomical regions were treated separately as depicted in Table 5. So as to bracket the historical experience of an initially high but declining fraction of tumors classified as squamous cell carcinomas of the upper bronchial tree calculations of initiation mutation rates used the limiting assumptions that 25, 50 or 75% of lethal tumors occurred in the upper tree. Assuming 25% of lethal tumors occur in the tracheal bronchial region increases this estimate by 1.3 fold, and assuming 75% decreases the estimate by 0.8 fold. As these assumptions had a relatively small effect on estimates of a first initiation mutation rate, estimates of 50% lethal tumors should be sufficient for the purposes of discerning any historical shift.

As may be seen in Table 5, the initiation parameter $2\,\tau^2\,\mathbf{r_i}\,\mathbf{r_j}$ was essentially identical for female nonsmokers, female smokers and male smokers for all birth decade cohorts analyzed. The average value is 2.9×10^{-11} with a range of 2.1 to 3.6×10^{-11} mutations^2/yr^2. Using a turnover rate estimate of $\tau = 5.7$ divisions per year and assuming *pro tempore* that the first and second initiation events have approximately the same rate, $\mathbf{r_i} = \mathbf{r_j}$, the rate of the first step of initiation, $\mathbf{r_i}$, would be about 7×10^{-7} events (mutations) per stem cell division, ten times higher than derived by observation of hprt mutations in peripheral T-cells of people from 0 to 75 years of age (Bigbee et al, 1998; Branda et al, 1993; Davies et al, 1992; Finette et al, 1994; Henderson et al, 1986; Hirai et al, 1995; Hou et al, 1995; Huttner et al, 1995; Liu et al, 1997; McGinniss et al, 1990; Tates et al, 1991) and ten times higher than the estimated first initiation step calculated for colon cancer (Table 5).

More importantly, if the necessary number of steps for initiation in the tracheal bronchial tree were the same for both smokers and nonsmokers, the hypothesis that cigarette smoke causes lung cancer by increasing initiation rates, either by increasing mutation rates or turnover rates, is untenable. Likewise, the possibility that smoking increases nuclear mutation rates but also creates an environment in which one less initiation mutation is required in smokers than in nonsmokers is excluded by the observation that the overall initiation parameters are identical. As 'n' decreases for a fixed initiation parameter, so would the geometric mean of the rate of the 'n' required initiation events, such that if fewer events were required in a smoker, initiation mutation rates in the smoker would be calculated to be less than in the nonsmoker. However, the converse possibility that smoking could increase the initiation mutation rate such that an initiation pathway in which one more event is required than in nonsmokers could now occur during a smoker's lifetime, is not excluded. The values of the geometric means of initiation mutation rates for the cases n= 1, 2, 3 or 4 do not create values outside the realm of information about lung cell mutation rates.

As seen in Table 5, initiation parameters in the peripheral bronchiolar region of the lung, $2\,\tau^2\,\mathbf{r_i}\,\mathbf{r_j}$, was essentially identical for female nonsmokers, female smokers and male smokers for all birth decade cohorts analyzed. The average value is 2.5×10^{-13} with a range of 1.8 to 3.6 x 10^{-13} mutations^2/yr^2 about two orders of magnitude less than that estimated above for the cells of the tracheal bronchial epithelium. This rate is essentially identical to that calculated for the colon.

If as these calculations suggest, the background rate of mutation is some ten times higher in the upper tracheal bronchial epithelium than in the lower bronchiolar epithelium, the

biological basis for this difference is unknown. A tenfold higher level of initiation mutation rate might nonetheless explain to some degree the observation known as "field cancerization," that multiple independent colonies of P53 mutants are found scattered in the head and neck, upper digestive tract, and tracheal bronchial areas (Chung et al, 1993; Tian et al, 1998; Sozzi et al, 1995).

The possibility of $\mathbf{n}$=1 must also be considered. For this case the initiation parameter is τ $\mathbf{r_i}$ and for the case of all cells in a turnover unit at risk results in an estimated value of $\mathbf{r_i} = 5.6 \times 10^{-12}$. If however only mutant stem cells could give rise to preneoplastic colonies, then this estimate would increase by a multiplicative factor equal to the number of cells in a turnover unit. Preliminary evidence (Li-Sucholeiki et al, unpublished; Coller et al, unpublished) suggests there are 32 cells in the turnover units of the human tracheal bronchial epithelium leading to an estimation that under the stem-cell-only assumption $\mathbf{r_i} = 1.8 \times 10^{-10}$. A rate of this magnitude might be imagined for a process in which a single very rare event, such as a particular single missense point mutation, would be required for initiation, e.g. activation of a proto-oncogene.

5.3.4 Historical Changes in Promotion Mutation Rates, $\mathbf{r_A}$.

For neither colon, lung nor any other form of mammalian cancer has any genetic event been found that is required for promotion of a preneoplastic cell into a cancerous cell. For simplicity then, promotion mutation rates for the case m = 1 required promotional events were calculated. No significant differences were found for average values of $\mathbf{r_A}$ among female nonsmokers (2.9×10^{-7}), female smokers (4.4×10^{-7}) and male smokers (3.0×10^{-7}). A similar

value for $\mathbf{r_A}$ for the case m=1 (2.0×10^{-7}), had previously been estimated for colon cancer. The three-stage carcinogenesis model does not predict an effect of cigarette smoking on promotion mutation rates.

5.3.5 Historical Changes in the Growth Rate of Precancerous Lesions, $(\alpha - \beta)$.

Precancerous lesions of the lung were found to have a growth rate of $(\alpha-\beta) = 0.17$ for female nonsmokers born in the mid 19th century. This same value would be obtained for nonsmoking males as shown in Figure 66. Net growth rates of precancerous lesions of nonsmokers are similar to those calculated for the colon and observed for the total mass of children (0.16) between the ages of 1.5 through puberty, suggesting the possibility that mutations permitting proliferative growth take the cell back to the growth rates of children.

Contrarily, $(\alpha-\beta) = 0.31$ for female smokers and 0.32 for male smokers. Almost doubling $(\alpha-\beta)$ has a major impact on the age-specific cancer rate in a population of persons at risk (Figure 52). The inverse of this parameter is the doubling time of the preneoplastic lesion in the two stage cancer model, so this change has the effect of changing the doubling time in nonsmokers of about 6 yrs to about 3 yrs in smokers. If promotional event rates were unaffected by smoking (Section 5.3.4) the smoking effect of doubling preneoplastic colony growth rates would be to reduce the period between initiation and promotion by about 40%, a difference of about twenty years.

5.3.6 High LOH and LOI levels in Human Lung Cancer

Considering (Table 6) the possibilities of m>1 perforce raise the estimated rates for $\mathbf{r_A}$. Obviously, as the number of events is increased, the mean geometric mutation rate must

increase. Because tumors show a very high fraction of loci with loss of heterozygosity or genomic imprinting, it is clear that the rates of these processes while low in normal tissues must be high in preneoplastic colonies.

Average LOH levels of all loci lung tumors of 0.6 (Wistuba, 1997, 1999), an estimated average time of 32 years between initiation and promotion for smokers (Δ_h – age of smoking adoption), and a division rate of preneoplastic lesions of 12.7 per year, lead to estimates of LOH rates (single outcome) in preneoplastic lesions to be approximately 7.4×10^{-4} events per year. Comparing this rate to the calculations of Table 6 leads to the observation that about 5 events at this rate could be accommodated in a promotional hypothesis involving independent losses of heterozygosity or genetic imprinting.

High rates of genomic instability have been noted from the time of Theodor Boveri (1914) as a characteristic of solid tumors and dysplastic lesions. It appears to be generally assumed that such genomic instability is a phenotypic requirement for neoplasia that is acquired after initiation, although no reported test of these assumptions exist.

Further speculation along these lines is not yet justified. Promotion may not involve genetic changes or even rare events of any kind. As noted above there are no known genes associated with tumor promotion in humans. An alternative mechanism is that individual cells in a growing preneoplastic colony become more isolated from normal cells by virtue of diminished intercellular electrochemical communication, as proposed by Bronk (1970). Here the idea is that intercellular communication in a tissue is essential for maintaining the order and balance represented by zero net growth during maintenance turnover in adult tissues. If initiation permits slow exponential growth by altering intracellular communication, then a natural consequence of

colony expansion would be isolation of cells from external signals. In this scenario, a cell embedded in a preneoplastic colony becomes electrochemically remote from normal cells of the tissue. In such a state it can either fail to receive important growth suppressing signals from afar or fail to disperse autocrine growth factors that it may make from time to time.

In this condition, m=0, and the principal term defining the expected number of promoted cells in a growing preneoplastic colony becomes $\alpha/(\alpha-\beta)\ (\alpha_c-\beta_c)/\alpha_c\ 2^{(\alpha-\beta)(t-a)}$. This has a value of approximately 75,000 cells when half of all preneoplastic lesions would be expected to have developed at least one surviving promoted cell at $(t - a) = \Delta_h$. Since potentially preneoplastic colonies such as "severely dysplastic lesions" or carcinoma *in situ* appear to attain such dimensions, Bronk's (1970) concept is encouraged by these calculations.

5.3.7 Conclusions

The assumption that cigarettes cause lung cancer by inducing genetic changes in normal cells, or that environmental factors affect cancer rates in other organs, must be recognized as an untested but critical hypothesis that drives a large fraction of public investment in environmental cancer research. As is the case for lung cancer and cigarettes, the clear record of environmental effects on cancer rates is generally assumed to be accounted by exposure of humans to environmental mutagens. Denissenko (1996) has found that the mutational spectrum of the TP53 gene created by benzo[a]pyrene, a putative carcinogen found in cigarette smoke, is the same as the mutational spectrum of the TP53 gene of cells taken from lung cancer samples, thereby suggesting a direct etiological link between cigarette smoke and lung cancer. Rodin and Rodin (2000) have however since found that the mutational spectrum of the TP53 gene in smokers' and nonsmokers' cells from lung cancer samples are not dissimilar.

Consistent with Rodin and Rodin's (2000) conclusions, Coller et al (1998) found that mitochondrial DNA mutant hotspots of a 100bp sequence in smokers' samples of normal bronchial tissue were not different from the mitochondrial DNA mutant hotspots found in nonsmokers' normal bronchial tissue samples. Elevated mutant fraction among smokers were also not observed. (Coller et al, 1998). Preliminary results have also revealed no significant differences in the mutant frequency of normal bronchial cells of three smokers and two nonsmokers of similar age for the mutant hotspots at base pair 746 and base pair 747 mutations of the TP53 gene. (Li-Sucholeiki et al, unpublished.)

The three-stage carcinogenesis model further predicts that:

1. All cigarette smokers are at lifetime risk of lung cancer risk, ($\mathbf{F_h} \sim 1.0$) but that only 0.1 of all nonsmokers are at lifetime risk.
1. The estimated net growth rate of preneoplastic lesions, $(\alpha-\beta)$, is ~0.32 for smokers and 0.17 for nonsmokers.
1. Estimates of initiation and promotion mutation rates are unaffected by smoking status.

These data form a new hypothesis: cigarettes cause lung cancer by creating a condition in all smokers by which certain undefined but naturally occurring preneoplastic lesions grow at a rate nearly twice that observed for preneoplastic lesions in nonsmokers. The quantitative differences observed for $(\alpha-\beta)$, growth rate of preneoplastic lesions, and $\mathbf{F_h}$, fraction of the population at lifetime risk, are sufficient to account for all of the differences between the age-specific lung cancer mortality rates between smokers and nonsmokers.

The hypothesis that smoking causes lung cancer by inducing genetic changes appears to be in error. Instead, one must consider the possibility that cigarette smoking selects for cells already carrying the necessary initiation mutations.

As a hypothetical example, the conditions for initiation in nonsmokers could involve mutational losses of both alleles of unknown gene X and in smokers, loss of both alleles of unknown gene Y. As the rates of loss of alleles for genes X and Y are similar and unaffected by smoking, the rate of creation of newly initiated cells is the same in smokers and nonsmokers. However, in nonsmokers the loss of gene Y creates a preneoplastic colony with net growth rate of 0.17 doublings per year and in smokers the loss of gene X a preneoplastic colony with net growth rate of 0.32. This parameter, $(\alpha-\beta)$, effects both the survival chances of a newly initiated cell and the growth rate of any surviving preneoplastic lesion.

A simplification of this model is X = Y wherein smoking provides directly or indirectly a growth stimulus for cells lacking functional gene X/Y. Indirect effects of external stimuli must be considered seriously. As previously reported, induction of mammary tumors in rats treated with methylnitrosourea selected previously occurring mutants in the H-ras gene rather than inducing the specific G→A transition as had been assumed. (Cha et al, 1994) The effect on the growth rates of initiated cells in the lung might occur as a result of chemical contact with the affected cells. It might also result from a more general effect on all cells of the lung epithelium as was subsequently found for methylnitrosourea on mammary cells. Even a systemic effect such as responses to central nervous system stimulation by nicotine cannot be eliminated by present evidence.

Genes now known to be involved in human initiation such as the APC gene in human colonic epithelium are involved in intercellular communication. It seems reasonable to propose

that the histological changes in the epithelium of smokers' lungs may also change the nature of intercellular communication. Changes in gene expression induced by cigarette smoking have also been reported relevant to DNA damage. (Willey et al, 1997; Crawford et al, 1998). These changes in turn may be permissive for preneoplasia if loss of unknown genes by normal mutational processes creates an initiated cell growing at the rate calculated for smokers' preneoplastic colonies.

As shown in Figure 72 the expected number of preneoplastic lesions in the upper bronchial tract at age 60 was about 0.39 so that 18 such persons would have been expected to display about 7 dysplastic lesions and carcinomas *in situ* if these were indeed all preneoplastic lesions. Wistuba et al (1997) reported 27 such lesions which is significantly greater than predicted by the model. They also reported that 7 of these 27 lesions were denominated as carcinoma *in situ*, a number similar to the predicted value of 7 for preneoplastic lesions. Interestingly only 35% of the male heavy smokers in Wistuba et al's (1997) study exhibited dysplasia in their upper bronchi [personal communication, Dr. Adi F. Gazdar]. Our model predicts that some 32% of smokers would carry such lesions based on an expected number of 0.39 per smoker (Figure 72) and a Poisson distribution of the number of expected lesions among smokers. Thus it is possible that the predicted values for preneoplastic lesions are consistent with those observed in smokers of age 60.

However, the apparent non-Poisson distribution of total moderate-severe dysplastic lesions plus carcinomas *in situ* is not anticipated by any version of the three-stage model of carcinogenesis. The model may be wrong, many lesions classified as dysplastic may not in fact be preneoplastic lesions, or dysplastic lesions may create multiple microcolonies in lungs as they grow. Finally, it may be that what are identified by pulmonary pathologists as carcinomas *in situ*

are in fact the only true preneoplastic lesions among a set of lesions identified as "dysplastic". Prediction and observation with regard to the number of preneoplastic lesions in the tracheal bronchial tracts of smokers may not be in disagreement.

Figure 73 compares the age-specific cumulative risk of developing lung cancer among smokers and former smokers reported by Peto et al (2000) to that predicted by the model assuming that population and physiologic risk parameters returned to those in a nonsmoking population soon after smoking cessation. The agreement between the observations and the prediction was satisfactory, assuming only that preneoplastic lesions of smokers grew at a doubling rate of 0.32, preexisting precancerous lesions in former smokers grew at a doubling rate of 0.17 soon after cessation, and that 10% of smokers after cessation were still at risk for new surviving preneoplastic lesions could arise.

When the alternative hypothesis that cigarette smoking might affect the rate of initiation (both initiation and promotion) was tested, the results were inconsistent with Peto et al's (2000) incidence data among former smokers.

Lastly, of great interest to lung pathologists has been the absolute rise of adenocarcinomas and small cell carcinomas as a function of history relative to the absolute number of squamous cell carcinomas (Figure 34). The three-stage carcinogenesis model predicts that the rise of adenocarcinomas is due to a decrease in the age of onset of the disease. However, as data was unavailable for ages < 40 years of age, the analysis could not distinguish whether this effect was due to an increase in promotion mutation rates in cells of the precursor lesions to adenocarcinomas, or alternately due to an increase in the growth rates of these precancerous lesions.

The prediction that cigarette smoking will be found to be without influence on the rates of nuclear genetic or other events required for initiation can be tested only by direct measurements in the upper and lower tracts of human lungs. The prediction that tracheal bronchial initiation (mutation) rates will be found to be considerably higher than peripheral bronchiolar rates when expressed as events per cell year requires similar studies in humans. Of additional interest is that there are no reports of point mutations induced in bacteria or human cells by cigarette smoke condensate or extracts in the literature of mutation research.

5.4 Alternate Hypotheses for the Observed Curvature of Age-Specific Mortality Rates

Recognition that the mortality/incidence curves by cancer reach a maximum and then decline, created the rationale for describing a subpopulation within each birthyear cohort that was at risk for that cancer. Necessarily, this construct also creates a subpopulation not at lifetime risk for cancer, such that mortality rates expectedly reach a maximum and decrease as a function of age since the subpoulation at risk dies off at a faster rate than the remaining population. As Cook et al (1969) have suggested, the possible existence of subpopulations at risk does not preclude other reasons that could likewise create the observed curvature of mortality data.

5.4.1 Assumption of Homogeneity in the Population for Secondary Risk Parameters r_i, r_A, ($\alpha - \beta$)

For simplification, the development of the mathematical models had assumed homogeneity among all the individuals who died of cancer such that stochastic distribution alone determined the age at which each individual died. Reasonably, inherited genetic and/or

environmental risk factors would affect the age of onset of cancer, such that the assumption of homogeneity in the population for cancer risk factors is not based on expectation.

Hemminki and Vaittinen (1998) have reported the elevated risk for breast cancer among Swedish mothers whose daughters have breast cancer (Figure 74). The relative risk of breast cancer for this set of mothers versus all mothers is constant at a value of 1.6 for all ages greater than 40 years.

The fact that the relative risk was a value greater than one and is constant for ages 42.5 to 92.5 suggests that the elevated risk of developing breast cancer among mothers whose daughters have breast cancer is due to a primary risk factor. If the primary risk factor were genetic, then a daughter with breast cancer is an obligatory carrier, such that the probability that her mother also carries the genetic risk factors is relatively higher than for a mother of a daughter who may not be a carrier of the genetic primary risk factor. (Since mothers and daughters likely share similar environments, the possibility that the primary risk factor is environmental cannot be excluded). Primary risk factors do not affect the age at which the mother is expected to develop breast cancer, such that the relative risk is expectedly constant.

Individuals inheriting a secondary risk factor (or exposed to an environmental secondary risk factor) are predisposed to get cancer at an earlier age, such that the relative risk would appear to decrease as a function of age based on probabilistic distribution. As the mother ages, the fact that she has not yet developed breast cancer, even though her daughter did, decreases the likelihood that she herself had inherited a secondary risk factor, such that the relative risk would appear to decrease as a function of age. Hemminki and Vaittinen's (1998) familial breast cancer data does reveal that for ages 27.5 to 42.5 there is an age-dependent decrease in the elevated risk of a mother, whose daughter has breast cancer, herself developing breast cancer. This suggests

that there may exist a secondary risk factor associated with early onset breast cancer. As the last of these mothers develop breast cancer, the remaining mothers who have inherited only the primary genetic risk factor are all at about the same risk of developing breast cancer, such that the calculated relative risk reaches a constant value. (Note that the daughters in this study had not yet reached age 60 so perforce some daughters who do have a secondary risk factor have not yet developed breast cancer. This limitation determines the age at which the relative risk is observed to reach a constant level; the decrease in relative risk may in fact extend to ages higher than 42.5, still decreasing to a level of 1.6, but more gradually).

The observation of a decreasing relative risk does reveal that there exists heterogeneity in the population at risk (breast cancer, Figure 74). Other cancers reveal similar results. Relative risk for colorectal cancer among parents whose children have colon cancer (data supplied by Prof. Hemminki, Karolinska Institute, Figure 74) reveals once again a constant relative risk among the elderly consistent with the hypothesis of a primary risk fraction for colon cancer. For the earlier age groups there is an elevated relative risk, consistent with a subpopulation having inherited genetic or sharing enviromental secondary risk factors.

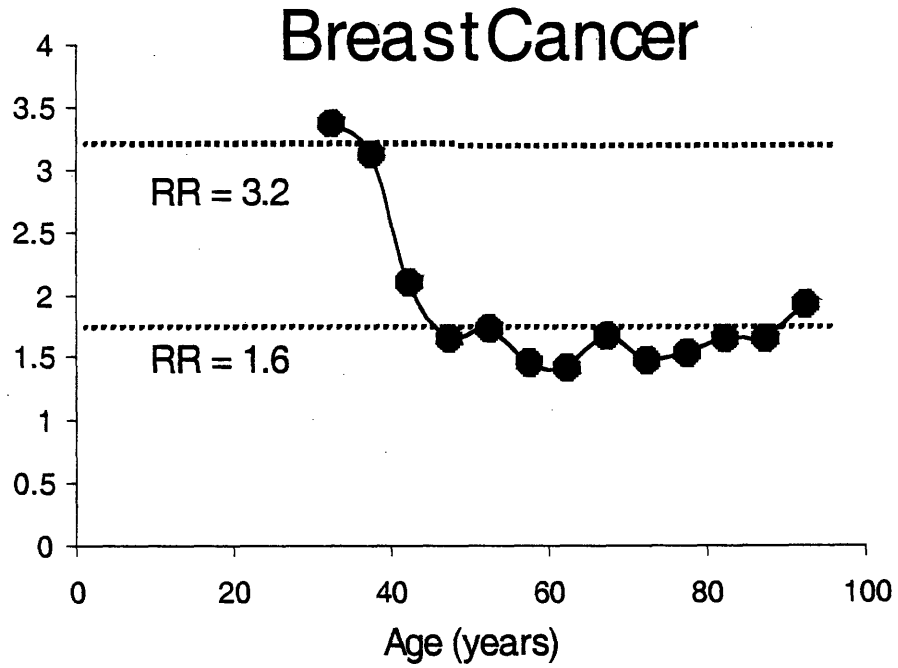
For the earliest ages, the relative risk is even higher consistent with the existence of FAP and HNPCC families for which familial inheritance of one of the two APC initiation mutations or a mutation in a mismatch repair gene respectively, elevates risk of developing colorectal cancer at an early age.

Fig. 74: Elevated risk of parent developing cancer given that a child has cancer

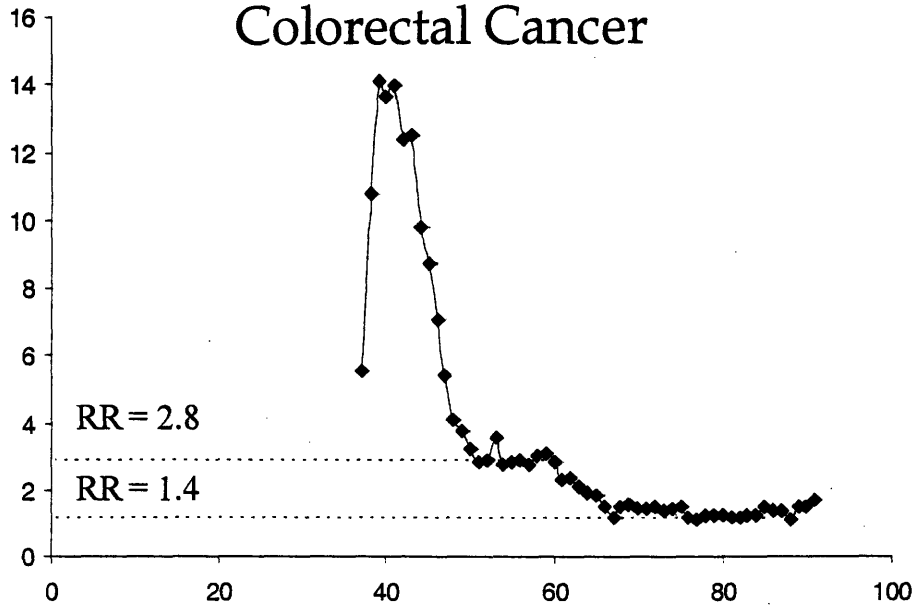
a) Relative risk of mothers, whose daughters have breast cancer, themselves developing breast cancer versus all mothers with daughters as a function of age

b) Relative risk of parents, who have a child with colorectal cancer, themselves developing colorectal cancer versus all parents as a function of age

a)


Breast Cancer
RR = 3.2
RR = 1.6
Age (years)
0
0.5
1
1.5
2
2.5
3
3.5
4
0
20
40
60
80
100

b)


Colorectal Cancer
RR = 2.8
RR = 1.4
0
2
4
6
8
10
12
14
16
0
20
40
60
80
100

5.4.2 Effect of heterogeneity in the population for secondary risk parameters $\mathbf{r_i}$, $\mathbf{r_A}$, $(\alpha - \beta)$, on the curvature of mortality data (n = 2)

In the two initiation mutation carcinogenesis model, the hazard function rises approximately linearly for physiologically observed rates of mutation ($< 10^{-5}$ per year). The slope of the linear rise is dependent on the mutation rate. Individuals who have inherited or due to environmental exposures have higher mutations rates expectedly would develop and die of cancer at an earlier age than individuals without these accelerating risk factors. The average mutation rate $\mathbf{r_i}$ of individuals who have not yet died would expectedly decrease as a function of age.

As a result, the slope of the hazard function would concurrently decrease thereby creating curvature on the observed mortality data. However, this would not actually cause the observed mortality curves to reach a maximum and then decrease unless the distribution of mutation rates were binodal, with distinct populations having high or low mutation rates. The only available data set for the distribution of mutation rates in humans is the mutant frequency distribution compiled by Dr. Aoy Tomita-Mitchell for the hprt locus of peripheral T-cells (Figure 8). Converting to mutation rates (Figure 75), the distribution is revealed to be approximately log-normal. Such a distribution would not create a condition by which the observed mortality would reach a maximum and decrease, if two initiation mutations were required. The slope of the mortality curve would decrease but the overall mortality curve would keep increasing, albeit at a gradually smaller rate.

Heterogeneity for the other two accelerating risk parameters, $\mathbf{r_A}$ and $(\alpha - \beta)$ would have similar effects if their distributions among all individual were like Figure 75. There is yet no

available data to determine actual distributions for these two parameters. However, in the case of $(\alpha - \beta)$, recall that distinct growth rates for smokers and nonsmokers were estimated (Table 5). Therefore the distribution of lung precancerous growth rates in a mixed population of smokers and nonsmokers based on these results would not resemble that of Figure 75 for mutation rates, but would rather be binodal with peaks at about 0.17 (nonsmokers) and 0.32 (smokers).

The three-stage carcinogenesis model had suggested that each cohort is comprised of three subpopulations: smokers all at risk for lung cancer, 10% of all individuals at risk for lung cancer by means other than smoking, and 90% of nonsmokers who were not at risk for lung cancer within a normal lifespan. This created a condition in which the mortality curves for cohorts of mixed populations of smokers and nonsmokers would reach a maximum and decrease, as the smoking population was depleted. The maximum in lung cancer mortality rates is observed due to the marked differences in the calculated growth rates of precancerous lesions between smokers and nonsmokers.

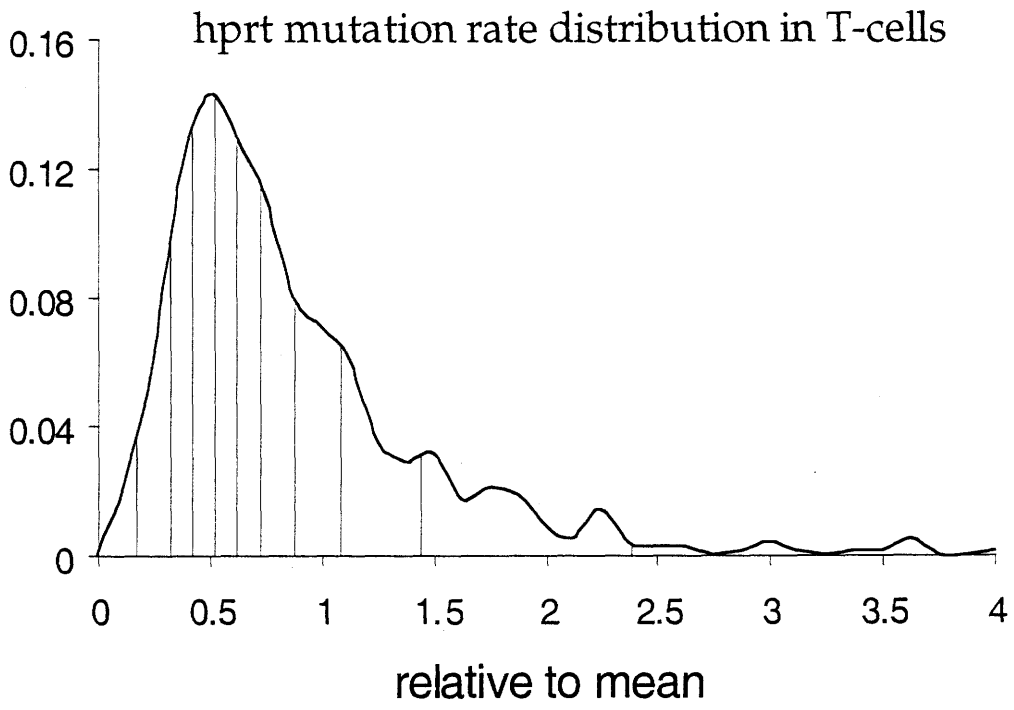
By the same reasoning that there is a decrease in the overall lung cancer mortality data after the smoking population is depleted due to their higher growth rate of precancerous lesions, it may be that the lung cancer mortality data for nonsmokers (Figure 66) also decreases if the 10% of nonsmokers found to be at risk for lung cancer are depleted more rapidly than the remaining nonsmokers (i.e. the other 90% of nonsmokers have a markedly lower precancerous growth rate). In this case, theoretically every individual would be at risk for developing cancer; the curvature was caused alone by a distribution of precancerous growth rates with three distinct populations: 0.32 (smokers), 0.17 (10% nonsmokers), and theoretically <0.08 (other 90% nonsmokers). However, this is mere speculation until the distribution of growth rates among all individuals is determined.

Fig. 75: Distribution of mutation rates of the hprt locus of peripheral T-cells

Mean mutation rate is 2.1×10^{-7} per cell year

Compiled by Dr. Aoy Tomita- Mitchell

(Bigbee, 1998; Branda, 1993; Davies, 1992; Finette, 1994; Henderson, 1986; Hirai, 1995; Hou, 1995; Huttner, 1995; Liu, et al, 1997; McGinniss et al, 1990; Tates et al, 1991)


hprt mutation rate distribution in T-cells
0.16
0.12
0.08
0.04
0
0
0.5
1
1.5
2
2.5
3
3.5
4
relative to mean

Heterogeneity of the three accelerating risk parameters (r_i, r_A, $\alpha - \beta$) could therefore create the observed maximum mortality data if and only if the distribution in these parameters showed subpopulations with distinct rates. This permits the possibility that all rather than some individuals are at risk for cancer. However, the data from Hemminki and Vaittinen (1998) does still show the existence of a distinct population at primary risk. (Section 5.4.1)

Evidently, both heterogeneity for secondary risk factors and the existence of a primary fraction at risk create curvature in the mortality data. Since the mortality data was initially analyzed assuming homogeneity, the observation of curvature was attributed entirely to the fraction at risk or to the factor, $\mathbf{f_h}$, correcting for death by connected forms of death. Correcting for the curvature caused by heterogenity predicts that our initial calculations may have underestimated the fraction at risk or the correction factor, $\mathbf{f_h}$. It is left for future research to answer this question.

5.4.3 Effect of heterogeneity in the population for secondary risk parameters $\mathbf{r_i}$, $\mathbf{r_A}$, $(\alpha - \beta)$, on the curvature of mortality data (n = 1)

Assuming homogeneity, the one initiation mutation carcinogenesis model predicts that the hazard function rises rapidly, plateaus and remains constant throughout the ages of a normal lifetime (assuming physiologically observed rates of mutation of $< 10^{-5}$ per year). The magnitude of this constant is directly proportional to the mutation rate. Again, individuals who have inherited or due to environmental exposures have higher mutations rates would develop and die of cancer at an earlier age than individuals without these accelerating risk factors. The

average mutation rate $\mathbf{r_i}$ of individuals who have not yet died would expectedly decrease as a function of age.

In the one initiation mutation model, since once the hazard function has 'plateaued' mortality is directly proportional to the mutation rate, the hazard function would concurrently decrease thereby creating not only curvature on the observed mortality data, but actually create the condition that observed mortality rates reach a maximum and decrease. Heterogeneity in the case of the one initiation mutation model could alone explain the observed maximum of mortality rates, and would make calculation of the fraction at risk not feasible.

The peaking of the observed mortality curve in the one initiation mutation model could also be predicted if exposure to a necessary environmental risk factor occurred only during part of life. This would create the condition in which the individuals who were exposed in this way would die at the elevated rate due to the environmental exposure. Once this fraction is killed off, the mortality rate would come down towards the level of unexposed individuals. Variability in the duration of exposure among all individuals would create a function that never plateaus, but rather reaches a maximum for the individuals exposed for the longest period and then continues to decrease thereafter.

We must however remember that as of yet no form of sporadic cancer has been shown to require but one initiation event, such that these hypothetical situations that would also create a mortality curve that peaks without regard of a fraction at risk may be of no consequence.

6. CONCLUSIONS

COLON CANCER:

42% of the population in the U.S. was calculated to be at risk for developing colon cancer within their lifetime, independent of gender or race.

No significant historical change in the calculated fraction at risk was found, suggesting that the primary risk factors for colon cancer are not environmental.

The calculated rate for the first initiation mutation in the colon was similar to the mutation rate observed for the hprt locus in human peripheral T-cells (~2.1 x 10^{-7} per cell year) and the spontaneous mutation rate of the hprt locus of human B-cells in culture.

Mutation rates were not found to increase significantly from the birthyear cohort of 1860 to the birthyear cohort of 1940. Environmental risk factors may play a secondary risk in colon cancer, albeit small, possibly by just increasing tissue renewal.

The calculated doubling rate of these lesions (~0.17-0.21) was found to be similar to the growth rate of children, suggesting that the required initiation events may have the net effect of reactivating pathways involved in child development.

The observed changes in colon cancer mortality rates can be explained alone by the increase in the lifespan of the American population during this century.

LUNG CANCER:

10% of nonsmokers and at least 94% of smokers were estimated to be at lifetime (primary) risk of death by lung cancer. The fraction at lifetime risk for all birthyear cohorts, consisting of mixed populations of smokers and nonsmokers, was found to be a simple linear function of reported cigarette use, independent of gender or race.

Growth rates of preneoplastic colonies of both genders and ethnic groups were found to be 0.17 and 0.32 doublings per year for nonsmokers and smokers respectively.

Rates of events such as genetic alterations necessary for initiation or promotion showed no differences among smokers and nonsmokers of either gender or ethnic group.

Incidence of lung cancer among former smokers can be predicted by a decrease in the growth rate of precancerous lesions of 0.32 doublings per year to 0.17 after smoking cessation.

Smoking causes lung cancer not by increasing the rate of genetic change in human lung epithelial cells, but by stimulating the growth of independently induced preneoplastic lesions in all smokers.

7. SUGGESTIONS FOR FUTURE RESEARCH

7.1 Verification of Conclusions

Mathematical models such as the one used herein are valuable in helping only to develop hypotheses, or to potentially disprove hypotheses. As such, mathematical models depend on actual *in vivo* or even *in vitro* experimentation to verify or prove any suggested hypotheses.

The conclusion that individuals are not placed at risk for colon cancer by exogenous exposures suggests that replication errors or endogenous damage are the leading causes of colon cancer.

Brindha Muniappan's work on the determination of the mutational spectrum of APC induced by DNA polymerase β should be extended to other potential endogenous processes (other replication pathways or repair mechanisms) and to other known tumor suppressor genes. Even if an exogenous agent could induce a mutational spectrum similar to that in a tumor suppressor gene of cells taken from cancer patients, it is imperative to have first discovered the background mutational hotspots so that one can distinguish between direct mutagenicity and selection for preeexisting mutants. The pathway by which an exogenous agent increases carcinogenicity would be of importance in directing potential research in therapies.

Dr. Li-Sucholeiki et al's work in determining mutant fraction of mutational hotspots of the TP53 gene should be continued so as to guarantee strong statistical significance for the current observation that smoking plays little, if any, role in the direct mutagenicity of lung epithelial cells. In the event that the tumor suppressor gene for lung cancer were to be discovered, of course the work should shift towards that gene. Even if work with further lungs were to show a somewhat elevated mutation rate in smokers' lungs, this observation would require the further verification that tissue renewal had not been accelerated in a smoker's lung.

For that purpose, Dr. Elena Gostjeva's work in determining mitotic rates in bronchial epithelia should be broadened to consider smokers and nonsmokers independently, not only of normal epithelia, but also of precancerous lesions.

Additionally, the increase in incidence of adenocarcinomas versus other forms of lung cancer is theoretically due to an increase in the growth rate of these lesions concurrently with a proportional increase in the mitotic rate (thereby keeping the term K_h constant as predicted by our model). This would be confirmed if mitotic rates were indeed different between smokers and nonsmokers.

Laslty, the current lung research has been conducted in the upper bronchial region. However, the models predict that mutation rates in the lower epithelia are 10 times lower, the same as calculated in the colon and in peripheral T-cells. This difference may be due to differential turnover rates or differential mutation rates. Alternately, the number of necessary initiation mutations is different for lung peripheral versus bronchial cells, such that the difference in calculated mutation rates may not be factual. For this purpose, the lung work should be extended to the lower peripheral portions.

7.2 Application of Model to Other Cancers

It would be of particular interest to determine whether or not the model predicts the association of cigarette smoke with other forms of cancer than lung. Peculiarly, no other cancer exhibits the 20-30 gender-dependent lag in the increase in occurence of cancer associated with the 20-30 year lag in females picking up the smoking habit. Paradoxically, there are many reports which show an elevated risk of other cancers among smokers (Doll and Hill, 1952; Hammond and Horn, 1958; Kahn, 1966).

Preliminary work in the analysis of pancreatic cancer, for which a 2-3 fold relative risk is reported among smokers, revealed that cigarettes were not a primary risk factor. However, the age of onset of pancreatic cancer has decreased in a similar way as have adenocarcinomas of the lung. Further analysis of this data should be done to determine whether or not smoking is a secondary risk factor for pancreatic cancer instead.

Of course, eventually every single form of cancer for which data exists should undergo similar analysis as has been done for colon and lung cancer.

7.3 Theoretical extension of carcinogenesis model – Effect of heterogeneity on estimated parameters

As explained in Section 5.4.2, heterogeneity in the population for secondary risk parameters creates curvature which had previously been wholly attributed to the potential existence of a primary fraction at risk. Excluding the existence of heterogeneity for secondary risk parameters would have the net effect of underestimating either the fraction at risk or the parameter correcting for the risk for connected forms of death. It would be recommended to ascertain the exact effect of this heterogeneity on the calculated population parameters.

Literature Cited

Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Brit J Cancer* 1954;8(1):1-12.

Armitage P, Doll R. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. *Brit J Cancer* 1957;9(2):161-169.

Auerbach O, Gere JB, Forman JB, Petrick TG, Smolin HJ, Muehsam GE, Kassouny DY, Stout AP. Changes in the bronchial epithelium in relation to smoking and cancer of the lung. *New Eng J Med* 1957;256(3):97-104.

Auerbach O, Stout AP, Hammond EC, Garfinkel LA. Bronchial epithelium oin former smokers. The *New Eng J Med* 1962;267(3):119-125.

Axtell LM, Asire AJ, Myers MH (eds), Cancer Patient Survival; Report Number 5, National Cancer Institute, Bethesda, MD, 1976.

Beart RW, Steele GD Jr., Menck JR, Chmiel JS, Ocwieja KE, Winchester DP. Management and survival of patients with adenocarcinoma of the colon and rectum: A national survey of the Commission on Cancer. *J Amer Coll Surg* 1995;181:225-236.

Benzer S, Freese E. Induction of specific mutations with 5-bromouracil. *Proc Natl Acad Sci USA* 1958;44:112.

Berg JW. Epidemiology of the different histologic types of lung cancer. In: Nettesheim P, Hanna Jr. MG, Deatherage Jr. JW, editors. Morphology of Experimental Respiratory Carcinogenesis. USAEC Division of Technical Information Extension, Oak Ridge, Tennessee, 1970:93-104.

Bigbee WL, Fuscoe JC, Grant SG, Jones IM, Gorvad AE, Harrington-Brock K, Strout CL, Thomas CB, Moore MM. Human in vivo somatic mutation measured at two loci: individuals with stably elevated background erythrocyte glycophorin A (gpa) variant frequencies exhibit normal T-lymphocyte hprt mutant frequencies. *Mut Res* 1998;397(2):119-136.

Bjerknes M, Cheng H, Hay K, Gallinger S. APC mutation and the crypt cycle in murine and human intestine. *Am J Pathol* 1997;150(3):833-839.

Branda RF, Sullivan LM, O'Neill JP, Falta MT, Nicklas JA, Hirsch B, Vacek PM, Albertini RJ. Measurement of HPRT mutant frequencies in T-lymphocytes from healthy human populations. *Mut Res* 1993;285(2):267-279.

Cariello NF, Keohavong P, Kat AG, Thilly WG. Molecular analysis of complex human cell populations: mutational spectra of MNNG and ICR-191. *Mut Res* 1990;231:165-176.

Cemerikic-Martinovic V, Trpinac D, Ercegovac M. Correlations between mitotic and apoptotic indices, number of interphase NORs, and histological grading in squamous cell lung cancer. *Micr Res Tech* 1998;40:408-417.

Cha RS, Thilly WG, Zarbl H. N-nitroso-N-methylurea-induced rat mammary tumors arise from cells with preexisting oncogenic Hras1 gene mutations. *Proc Natl Acad Sci U S A* 1994;;91(9):3749-3753.

Chen J, Thilly WG. Mutational spectra vary with exposure conditions: benzo[a]pyrene in human cells. *Mut Res* 1996;357(1-2):209-217.

Cloutier J-F, Drouin R, Castonguay A. Treatment of human cells with N-nitroso(acetoxymethyl)methylamine: Distribution patterns of piperidine-sensitive DNA damage at the nucleotide level of resolution are related to the sequence context. *Chem Res Toxicol* 1999;12:840-849.

Coller HA, Khrapko K, Torres A, Frampton MW, Utell MJ, Thilly WG. Mutational spectra of a 100-base pair mitochondrial DNA target sequence in bronchial epithelial cells: a comparison of smoking and nonsmoking twins. *Cancer Res* 1998;58(6):1268-1277.

Cook PJ, Doll R, Fillingham SA. A mathematical model for the age distribution of cancer in man. *Int J Cancer* 1969;4(1):93-112.

Cooper DA, Crane AR, Boucot KR. Primary carcinoma of the lung in nonsmokers. *Arch Environ Health* 1968;16:398-400.

Coulondre C, Miller JH. Genetic studies of the lac repressor. IV. Mutagenic specificity in the *lacI* gene of *Escherichia coli. J Mol Biol* 1977;117:577-606.

Cui H, Horon IL, Ohlsson R, Hamilton SR, Feinberg AP. Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability. *Nat Medicine* 1998;4(11):1276-1280.

Cutler SJ, Ederer F., editors, End Results and Mortality Trends in Cancer; NCI Monograph Number 6. National Cancer Institute, Bethesda, MD, 1961.

Davies MJ, Lovell DP, Anderson D. Thioguanine-resistant mutant frequency in T-lymphocytes from a healthy human population. *Mut Res* 1992;265(2):165-171.

Dennisenko MF, Pao A, Tang M, Pfeifer GP. Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science* 1996;274(5286):430-432.

de Nooij-van Dalen AG, van Buuren-van Seggelen VH, Lohman PH, Giphart-Gassler M, Chromosome loss with concomitant duplication and recombination both contribute most to loss of heterozygosity in vitro. *Genes Chromosomes Cancer* 1998;21(1):30-38.

Dewanji A, Moolgavkar SH, Luebeck EG. Two-mutation model for carcinogenesis: Joint analysis of premalignant and malignant lesions. *Mathematical Biosciences* 1991;104(1):97-109.

Dewanji A, Venzon DJ, Moolgavkar SH. A stochastic two-stage model for cancer risk assessment. II. The number and size of premalignant clones. *Risk Anal.* 1989;9(2):179-187.
Doll R, Hill AB. A study of the aetiology of carcinoma of the lung. *Br Med J* 1952;2:1271-1286.

Doll, R. The age distribution of cancer in man, in Cancer and Aging, Nordiska Bokhandelns Förlag, Stockholm, 1968:15-40.

Eisenberg H, Sullivan PD, Connelly RR. Cancer in Connecticut: Survival experience, 1935-1962. Connecticut State Department of Health, Hartford, CT, 1968.

Enstrom JE, Godley FH. Cancer mortality among a representative sample of nonsmokers in the United States during 1966-68. *JNCI* 1980;65(5):1175-1183.

Ernster VL. Epidemiology of lung cancer in women. *Ann Epidemiol* 1994;4:102-110.

Finette BA, Sullivan LM, O'Neill JP, Nicklas JA, Vacek PM, Albertini RJ. Determination of hprt mutant frequencies in T-lymphocytes from a healthy pediatric population: statistical comparison between newborn, children and adult mutant frequencies, cloning efficiency and age. *Mut Res* 1994;308(2):223-231.

Friend SH, Horowitz JM, Gerber MR, Wang X-F, Bogenmann E, Li FP, Weinberg RA. Deletions of a DNA sequence in retinoblastomas and mesenchymal tumors: organization of the sequence and its encoded protein. *Proc Nat Acad Sci* 1987;84:9059-9063.

Fuller CE, Davies RP, Williams GT, Williams ED. Crypt restricted heterogeneity of goblet cell mucus glycoprotein in histologically normal human colonic mucosa: a potential marker of somatic mutation. *Br J Cancer* 1990;61:382-384.

Gailani MR, Stahle-Backdahl M, Leffell DJ, Glynn M, Zaphiropoulos PG, Pressman C, Unden AB, Dean M, Brash DE, Bale AE, Toftgard R. The role of the human homologue of Drosophila patched in sporadic basal cell carcinomas. *Nature Genet* 1996;14:78-81.

Garcia-Patiño E, Gomendio B, Lleonart M, Silva JM, Garcia JM, Provencio M, Cubedo R, España P, Ramón y Cajal S, Bonilla F. Loss of heterozygosity in the region including the BRCA1 gene on 17q in colon cancer. *Cancer Genet Cytogenet* 1998;104:119-123.

Gennett IN, Thilly WG. Mapping large spontaneous deletion endpoints in the human HPRT gene. *Mut. Res*1988;201(1):149-160.

Giovannucci E, Goldin B. The role of fat, fatty acids, and total energy intake in the etiology of human colon cancer. *Am J Clin Nutr* 1997;66(6 Suppl):1564S-1571S.

Gnarra JR, Tory K, Weng Y, Schmidt L, Wei MH, Li H, Latif F, Liu S, Chen F, Duh FM, et al. Mutations of the VHL tumour suppressor gene in renal carcinoma. *Nat Genet* 1994;7(1):85-90.

Grist SA, McCarron M, Kutlaca, A, Turner DR, Morley AA. *In vivo* human somatic mutation: Frequency and spectrum with age. *Mut Res* 1992;266(2):189-196.

Hamill PV, Drizd TA, Johnson CL, Reed RB, Roche AF, Moore WM. Physical Growth: National Center for Health Statistics percentiles. *Am J Clin Nutr* 1979;32(3):607-629.

Hammond EC, Horn D. Smoking and death rates: Report on forty-four months of follow-up of 187,783 men. *JAMA* 1958;166:1159-72,1294-1308.

Harris JE. Cigarette smoking among successive birth cohorts of men and women in the United States during 1900-80. *JNCI* 1983;71(3):474-478.

Henderson L, Cole H, Cole J, James SE, Green M. Detection of somatic mutations in man: evaluation of the microtitre cloning assay for T-lymphocytes. *Mutagenesis* 1986;1(3):195-200.

Hernandez-Boussard TM, Hainaut P. A specific spectrum of p53 mutations in lung cancer from smokers: review of mutations compiled in the IARC p53 database. *Environ Health Perspect* 1998;106(7):385-391.

Herrero P, Thilly WG, Morgenthaler S. A Stochastic Model of Carcinogenesis. In: Fernholz LT, Morgenthaler S, Stahel W, editors. Statistics in Genetics and in the Environmental Sciences. Birkhaeuser Verlag, Basel, 2000:77-88.

Heston JF, Kelly JB, Meigs JW, Flannery JT, Cusano MM, Young JL Jr. (eds). Forty-five years of Cancer Incidence in Connecticut: 1935-79, National Cancer Institute, Bethesda, MD,1986.

Hirai Y, Kusunoki Y, Kyoizumi S, Awa AA, Pawel DJ, Nakamura N, Akiyama M. Mutant frequency at the HPRT locus in peripheral blood T-lymphocytes of atomic bomb survivors. *Mut Res* 1995;329(2):183-196.

Hou SM, Falt S, Steen AM. Hprt mutant frequency and GSTM1 genotype in non-smoking healthy individuals. *Environ Mol Mutagen* 1995;25(2):97-105.

Huttner E, Holzapfel B, Kropf S. Frequency of HPRT mutant lymphocytes in a human control population as determined by the T-cell cloning procedure. *Mut Res* 1995;348(2):83-91.

Jass JR, Edgar S. Unicryptal loss of heterozygosity in hereditary non-polyposis colorectal cancer. *Pathology* 1994;26:414-417.

Jen, J, Powell SM, Papadopoulos N, Smith KJ, Hamilton SR, Vogelstein B, Kinzler KW. Molecular determinants of dysplasia in colorectal lesions. *Cancer Research* 1994;54(21):5523-5526.

Kahn HA. The Dorn study of smoking and mortality among U.S. veterans: Report on eight and one-half years of observation. In: Haenszel W, editor. Epidemiological Approaches to the Study of Cancer and Other Chronic Diseases, National Cancer Institute Monograph 19, Bethesda, Maryland, 1966:1-126.

Kamb A, Shattuck-Eidens D, Eeles R, Liu Q, Gruis NA, Ding W, Hussey C, Tran T, Miki Y, Weaver-Feldhaus J, McClure M, Aitken JF, Anderson DE, Bergman W, Frants R, Goldgar DE, Green A, MacLennan R, Martin NG, Meyer LJ, Youl P, Zone JJ, Skolnick MH, Cannon-Albright LA. Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nature Genet* 1994;8:22-26.

Keohavong P, Thilly WG. Mutational spectrometry: A general approach for hot-spot point mutations in selectable genes. *Proc Natl Acad Sci USA* 1992;89:4623-4627.

Kennaway EL. Cancer of the penis and circumcision in relation to the incubation period of cancer. *Brit J Cancer* 1947;1(4):335.

Kermack WO, McKendrick AG, McKinlay PL. Death-rates in Great Britain and Sweden: Expression of specific mortality rates as products of two factors, and some consequences thereof. *Journ. Of Hyg* 1934;34(4):433-457.

Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, Smith KJ, Preisinger AC, Hedge P, McKechnie D. Identification of FAP locus genes from chromosome 5q21. *Science* 1991;253(5020):661-665.

Kinzler WW, Vogelstein B. Lessons from hereditary colon cancer. *Cell* 1996;87:159-170.

Knudson AG Jr. Mutation and cancer: Statistical study of retinoblastoma. *Proc Natl Acad Sci USA* 1971;68(4):820-823.

Kohno H, Hiroshima K, Toyozaki T, Fujisawa T, Ohwada H. p53 mutation and allelic loss of chromosome 3p, 9p of preneoplastic lesions in patients with nonsmall cell lung carcinoma. *Cancer* 1999;85(2):341-347.

Kuhn III C. Normal anatomy and histology. In: Thurlbeck WM, Churg AM, editors. Pathology of the Lung, 2nd Edition. Thieme Medical Publishers, New York, 1995:1-36.

Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystrom-Lahti M. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* 1993;75(6):1215-1225.

Levy D, Smith KJ, Beazer-Barclay Y, Hamilton SR, Vogelstein B, Kinzler KW. Inactivation of both APC alleles in human and mouse tumors. *Cancer Research* 1994;54(22):5953-5958.

Liu Y, Cortopassi G, Steingrimsdottir H, Waugh AP, Beare DM, Green MH, Robinson DR, Cole J. Correlated mutagenesis of bcl2 and hprt loci in blood lymphocytes. *Environ Mol Mutagen* 1997;29(1):36-45.

Loechler EL. Adduct-induced base-shifts: a mechanism by which the adducts of bulky carcinogens might induce mutations. *Biopolymers* 1989;28(5):909-927

McGinniss MJ, Falta MT, Sullivan LM, Albertini RJ. *In vivo* hprt mutant frequencies in T-cells of normal human newborns. *Mut Res* 1990;240(2):117-126.

Miller BA, Ries LAG, Hankey BF, Kosary CL, Harras A, Devesa SS, Edwards BK, editors, SEER Cancer Statistics Review: 1973-1990. National Cancer Institute, Bethesda, MD, 1993.

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

Moolgavkar SH, Knudson AG Jr. Mutation and cancer: A model for human carcinogenesis. *J Natl Cancer Inst* 1981;66(6):1037-1052.

Moolgavkar SH, Dewanji A, Venzon DJ. A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. *Risk Anal* 1988;8(3):383-392.

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

Moolgavkar SH, Luebeck EG, de Gunst M, Port RE, Schwarz M. Quantitative analysis of enzyme-altered foci in rat hepatocarcinogenesis experiment: I. Single-agent regimen. *Carcinogenesis* 1990;11(8):1271-1278.

Moolgavkar SH, Luebeck EG Multistage carcinogenesis: Population-based model for colon cancer. *J Natl Cancer Inst* 1992;84(8):610-618.

Nesnow S, Ross JA, Stoner GD, Mass MJ. Mechanistic linkage between DNA adducts, mutations in oncogenes and tumorigenesis of carcinogenic environmental polycyclic aromatic hydrocarbons in strain A/J mice, *Toxicology* 1995;105(2-3):403-413.

Nordling CO. A new theory on the cancer-inducing mechanism. *Brit J Cancer* 1953;7:68-72.

Oller AR, Thilly WG. Mutational spectra in human B-cells. Spontaneous, oxygen and hydrogen peroxide-induced mutations at the hprt gene. *J Mol Biol*1992;228(3):813-826.

Peterson LA, Hecht SS. O6-methylguanine is a critical determinant of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone tumorigenesis in A/J mouse lung. *Cancer Res* 1991;51(20):5557-5564.

Peto R, Roe FJ, Lee PN, Levy L, Clack J. Cancer and ageing in mice and men. *Br J Cancer* 1975;32(4):411-426.

Peto R. "Epidemiology, Multistage Models, and Short-term Mutagenicity Tests", Origins of Human Cancer, Cold Spring Harbor Laboratory Cold Spring Harbor, NY, 1977:1403-1429.

Peto R, Lopez A, Boreham J, Thun M, Heath C. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. Oxford University Press, Oxford, UK, 1988.

Peto R, Lopez A, Boreham J, Thun M, Heath C. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet* 1992;339:1268-1278.

Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case-control studies. *BMJ* 2000;321(7257):323-329.

Platt R. [letter to the editor] *Lancet* 1955;i:867.

Powell SM, Zilz N, Beazer-Barclay Y, Bryan TM, Hamilton SR, Thibodeau SN, Vogelstein B, Kinzler KW. APC mutations occur early during colorectal tumorigenesis. *Nature* 1992;359(6392):235-237.

Prodi V, Mularoni A. Particulate deposition in smoking. In: Cumming G, Bonsignore G, editors. Smoking and the lung. Plenum Press, New York, 1984:249-285.

Ragnarsson G, Eiriksdottir G, Johannsdottir JTh, Jonasson JG, Egilsson V, Ingvarsson S. Loss of heterozygosity at chromosome 1p in different solid tumours: association with survival. *Br J Cancer* 1999;79(9/10):1468-1474.

Resta N, Simone C, Mareni C, Montera M, Gentile M, Susca F, Gristina R, Pozzi S, Bertario L, Bufo P, Carlomagno N, Ingrosso M, Rossini FP, Tenconi R, Guanti G. STK11 mutations in Peutz-Jeghers syndrome and sporadic colon cancer. *Cancer Research* 1998;58:4799-4801.

Ries LAG, Pollack ES, Young JL Jr. Cancer Patient Survival: Surveillance, Epidemiology, and End Results Program, 1973-79. *JNCI* 1983;70(4):693-707.

Ries LAG, Kosary CL, Hankey BF, Miller BA, Harras A, Edwards BK, editors, SEER Cancer Statistics Review: 1973-1994. National Cancer Institute, Bethesda, MD, 1997.

Ries LAG, Kosary CL, Hankey BF, Miller BA, Clegg L, Edwards BK, editors, SEER Cancer Statistics Review: 1973-1996. National Cancer Institute, Bethesda, MD, 1999.

Rodin SN, Rodin AS. Human lung cancer and p53: the interplay between mutagenesis and selection. *Proc Natl Acad Sci USA* 2000;97(22):12244-12249.

Rouleau GA, Merel P, Lutchman M, Sanson M, Zucman J, Marineau C, Hoang-Xuan K, Demczuk S, Desmaze C, Plougastel B, Pulst SM, Lenoir G, Bijlsma E, Fashold R, Dumanski J, de Jong P, Parry D, Eldrige R, Aurias A, Delattre O, Thomas G. Alteration in a new gene encoding a putative membrane-organizing protein causes neuro-fibromatosis type 2: *Nature* 1993;363:515-521.

Staunton MJ, Gaffney EF. Tumor type is a determinant of susceptibility to apoptosis. *Am J Clin Pathol* 1995;103:300-307.

Tates AD, van Dam FJ, van Mossel H, Shoemaker H, Thijssen JC, Woldring VM, Zwinderman AH, Natarajan AT. Use of the clonal assay for the measurement of frequencies of HPRT mutants in T-lymphocytes from five control populations. *Mut Res* 1991;253(2):199-213.

Thun MJ, Lally CA, Flannery JT, Calle EE, Flanders WD, Heath Jr. CW. Cigarette smoking and changes in the histopathology of lung cancer. *JNCI* 1997;89(21):1580-1686.

Tormanen U, Nuorva K, Soini Y, Paakko P. Apoptotic activity is increased in parallel with the metaplasia-dysplasia-carcinoma sequence of the bronchial epithelium. *Br J Cancer* 1999;79(5/6):996-1002

Uhrhammer N, Bay J, Pernin D, Rio P, Grancho M, Kwiatkowski F, Gosse-Brun S, Daver A, Bignon Y. Loss of heterozygosity at the ATM locus in colorectal carcinoma. *Oncol Rep* 1999;6(3):655-658.

U.S. Bureau of the Census. Mortality Statistics, Special Reports (1930-1936). Washington Government Printing Office.

U.S. Department of Health and Human Services. Vital Statistics of the United States, Volume II-Mortality Part A (1937-1997). U.S. Government Printing Office, Hyattsvile, Maryland.

U.S. Department of Health and Human Services. In: Cancer Mortality in the United States: 1950-1977. U.S. Government Printing Office, Hyattsvile, Maryland, 1982. Appendix III: Percy C, Stanek E III, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics, 467-475.

Vincent TN, Satterfield JV, Ackerman LV. Carcinoma of the lung in women. *Cancer* 1965;18:559-570.

Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, Bos JL. Genetic alterations during colorectal-tumor development. *N Engl J Med* 1988;319(9):525-532.

Vogelstein B, Fearon ER, Kern SE, Hamilton SR, Preisinger AC, Nakamura Y, White R. Allelotype of Colorectal Carcinomas. *Science* 1989;244:207-211.

Walter JB, Pryce DM. The histology of lung cancer. *Thorax* 1955;10:107-116.

Walter JB, Pryce DM. The site of origin of lung cancer and its relation to histological type. *Thorax* 1955;10:117-126.

Weinstein WM, Tygat GN, Chen M, Band PR. In vivo studies of cell proliferation and kinetics in the human jejunal mucosa. *Gastroenterology* 1973;64:A137/820.

Wistuba II, Lam S, Behrens C, Virmani AK, Fong, KM, LeRiche J, Samet JM, Srivastava S, Minna JD, Gazdar AF. Molecular damage in the bronchial epithelium of current and former smokers. *J Nat Cancer Inst* 1997;89(18):1366-1373.

Wistuba II, Behrens C, Milchgrub S, Bryant D, Hung J, Minna JD, Gazdar AF. Sequential molecular abnormalities are involved in the multistage development of squamous cell lung carcinoma. *Oncogene* 1999;18:643-650.

Wistuba II, Behrens C, Virmani AK, Milchgrub S, Syed S, Lam S, Mackay B, Minna JD, Gazdar AF. Allelic losses at chromosome 8p21-23 are early and frequent events in the pathogenesis of lung cancer. *Cancer Research* 1999;59:1973-1979.

Wright N, Watson A, Morley A, Appleton D, Marks J, Douglas A. The cell cycle time in the flat (aviollous) mucosa of the human small intestine. *Gut* 1973;14:603-606.

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

Wynder EL, Berg JW. Cancer of the lung among nonsmokers. *Cancer* 1967;20:1161-1172.

Zheng T, Holford TR, Boyle P, Chen Y, Ward BA, Flannery J, Mayne ST. Time trend and the age-period-cohort effect on the incidence of histologic types of lung cancer in Connecticut, 1960-1989. *Cancer* 1994: 74(5):1556-1567.