Lung cancer mortality prediction using multi-state population smoking models

M. S. Clements

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

University of Sydney

2001

Declaration

I hereby declare that the work presented in this thesis, except where noted otherwise, is solely my own work and that to the best of my knowledge the work is original except where otherwise indicated by reference to other authors. No part of this work has been submitted for any other degree or diploma.

The report to the Commonwealth, provided in Appendix A, was written jointly with Associate Professor Richard Taylor. We jointly developed the analytical approach. I performed all of the analysis and completed the first draft of the report. Richard re-drafted the report, particularly the introduction and the discussion of lung cancer models.

Mark Clements
December 2001

Abstract

Aims

To investigate methods to predict lung cancer mortality by using multi-state population-based smoking models and vital statistics. To apply the methods to Australian populations.

Background

The main predictor of lung cancer mortality is smoking, however smoking patterns in Australia changed rapidly throughout the twentieth century.

Methods

A multi-stage smoking model included states for never, current and former smokers, with transitions for uptake, cessation, recall error and differential mortality. Methods for estimation of transition intensities were developed for three different sampling frames: retrospective, intercensal and multiple cross-sectional surveys. Statistical methods included non-linear estimation for mathematical models and local likelihood estimation. Representative all cause mortality rate ratios were estimated by combining aetiological data with Australian descriptive data. Multi-state models were back-projected to adjust for differential survival and recall error, and then forward-projected to predict smoking, prevalence dose and duration. Multi-stage lung cancer regression models were fitted with the smoking parameters. Finally, projected exposure parameters were applied to the mortality rate equations to estimate lung cancer mortality projections

Results

Smoking initiation and cessation rates were not proportional over time, so that non-parametric hazard estimation was required. Recall error from former to never smokers was estimated at 1–2% per annum of former smokers. Smoking dose changed appreciably over time, peaking during 1960–1975. Moreover, age-specific smoking duration increased among the younger female birth cohorts. Lung cancer mortality rates for males were predicted to continue to decline, however the level at 2028 was sensitive to future cessation rates. Moreover, the slope for the decline of female rates was sensitive to cessation rates for future decades.

Conclusions

Limitations include: imprecise estimates for cessation and recall error rates; validity of retrospective estimates; and no adjustment for other aspects of smoking exposure such as cigarette composition and construction. Birth cohorts were restricted to those born after 1920, which has improved validity and reduced precision. A multi-disciplinary approach to modelling, combining epidemiology, mathematical modelling and statistics provides insights into lung cancer mortality. The main public health implication is that continued smoking cessation efforts are required to ensure the decline of lung cancer rates in the future.

Acknowledgements

I am grateful to Professor Bruce Armstrong for his supervision, support and encouragement throughout this doctorate. He has offered a sense of what can be achieved through public health and epidemiology. I thank Associate Professor Ian Marschner for his willing support and supervision. I gratefully acknowledge the encouragement and guidance from Marylon Coates. I have been fortunate to share an office with her. I wish to thank Bruce, Ian, Marylon and Claire Vajdic for reading draft manuscripts.

I gratefully acknowledge the following for intellectual input:

- Ed Mitchell and Murray Laugesen for introducing me to epidemiology and tobacco control, respectively
- Alistair Woodward for the initial idea
- Phil Weinstein for creativity
- Bob Gibberd and Annette Dobson for direction
- Claire Vajdic, David Smith, Dianne O'Connell and Charlie Blumer for thoughtful conversations.

I am grateful for financial support from the Chancellors Committee at the University of Sydney and for support from the Cancer Council NSW. For provision of data, I acknowledge the New Zealand Ministry of Health, Jim Fraser and the New Zealand Health Information Service, Statistics New Zealand, the Australian Bureau of Statistics, and the Australian Social Sciences Data Archive. I am grateful for the collegiality from those in the Department of Public Health and Community Medicine at the University of Sydney and those at the Cancer Council. The Cancer Council has offered excellent hospitality.

My friends, family and Nina, my partner, have been very supportive throughout the doctorate. I could not ask for better people.

Publications

Clements, M., Armstrong, B. and Marschner, I. (2001). Advantages to using multi-state models for estimating population smoking exposure. *Am J Epidemiol* 153(11):451 Suppl. S.

Clements, M., Armstrong, B. and Marschner, I. (2001). Will US smoking prevalence "inexorably continue to decline"? *Am J Epidemiol* 153(11):472 Suppl. S. [Student poster prize at 2001 Epidemiology Congress]

Contents

List of Tables

List of Figures

Table of Abbreviations

Abbreviation	Description
ACCV	Anti-Cancer Council of Victoria
AIC	Akaike's Information Criterion
CPS-I	Cancer Prevention Study I
CPS-II	Cancer Prevention Study II
GAM	Generalised additive model
GLM	Generalised linear model
ICD	International Classification of Diseases
NDSHS	National Drugs Strategy Household Survey
NHIS	National Health Interview Survey
NMFS	National Mortality Followback Study
NZ	New Zealand
RFPS	Risk Factor Prevalence Study
WHO	World Health Organisation

Table of Symbols

Symbol(s)	Description
N, C, X, E, T	Never, current, former and ever smokers, and total population
n, c, x, e	Never, current, former and ever smokers (subscripts)
π	Prevalence
μ, μ_0	Total mortality rate, never smoker mortality rate
RR	Rate ratio
S	Survival
α_I	Initiation rate
α_Q	Cessation rate
α_E	Recall error rate by former smokers as never smokers
α_{Mig}	Migration rate
P_{jk}	Transition probability from state j to state k
$\boldsymbol{P}$	Transition probability matrix
α_{jk}	Transition intensity from state j to state k
$\boldsymbol{\alpha}$	Transition intensity matrix
A_{jk}	Cumulative intensity from state j to state k
$\boldsymbol{A}$	Cumulative intensity matrix
Dur	Duration
Dose	Dose per smoker
CumDose	Cumulative dose per smoker
Cons	Consumption *per caput* of population
CumCons	Cumulative consumption *per caput* of population
$\text{Dur}^*, \text{Dose}^*$	Duration and dose for (age-5 years)
λ	Lung cancer mortality rate
λ_0	Lung cancer mortality rate for never smokers

Chapter 1

Introduction

Abstract

Tobacco smoking is a major public health issue, for which lung cancer is a marker disease. Throughout the twentieth century, there were rapid changes in lung cancer mortality and smoking exposure in Australia and New Zealand. Cumulative measures of smoking exposure are important aetiological parameters for lung cancer mortality. For health priority setting, these cumulative measures are required to predict the impact of tobacco control interventions on lung cancer mortality. Moreover, predictions for lung cancer mortality can potentially be improved by the use of such measures. This thesis aims to develop multi-state population-based smoking models to improve exposure estimation, and to apply those methods to predict lung cancer mortality in Australia.

1.1 Background

1.1.1 Tobacco as a public health issue

For Australia in 1996, almost 17,000 deaths and 10 percent of the total burden of disease were attributed to tobacco smoking (Mathers et al., 1999). A number of diseases contributed to the total burden, including cardiovascular disease, cerebrovascular disease and lung cancer. The large burden of disease due to tobacco has also been recognised as a global problem (Murray and Lopez, 1997b).

1.1.2 Lung cancer

Lung cancer is an important marker disease for tobacco-related burden of disease. The rise of lung cancer parallels the rise of understanding of tobacco as a public health issue. The first journal article published on the causal association between smoking and lung cancer was in Germany in 1939 (Smith and Ebrahim, 2001). Later, British and American case-control studies and cohort studies in the 1950s and 1960s put tobacco on to the public health agenda for developed countries (U.S. Department of Health Education and Welfare, 1979; Peto et al., 1994).

There is a close link between changes in tobacco exposure and changes in lung cancer incidence and mortality, expressed by a large population attributable fraction. For Australia in 1992, English and colleagues estimated the population attributable fraction of lung cancer deaths to tobacco smoking as 85% for males and 77% for females (English et al., 1995). To paraphrase, if no one had smoked then over three quarters of lung cancer deaths would potentially have been avoided.

This close link has two important implications. First, the rise in lung cancer can serve as an historical summary of a population's cumulative exposure to tobacco smoking. Second, changes in smoking exposure can be used to predict future lung cancer incidence and mortality. Note that lung cancer is a limited measure for assessing tobacco interventions, primarily because of the time lag between change of exposure and any change of lung cancer risk.

Lung cancer is highly fatal, where survival to five years after cancer registration is typically less than 15% (Supramaniam et al., 1999; Berrino et al., 1999). One consequence is that patterns for incidence and mortality closely mirror each other. Australia has limited national data available on lung cancer incidence, so that lung cancer mortality provides a better measure for longer-term time series. Use of mortality has the advantage of measuring a specific dimension of disease burden, but death may be a combination of environmental exposures, related to incidence, and factors associated with early detection and treatment, related to survival.

In the following sections, a brief review of the descriptive epidemiology of lung cancer is given. The intention is to provide a setting for the thesis. Data sources are detailed in Chapter 2. Methods used to smooth the age-specific rates and prevalence are outlined in Chapter 3.

1.1.2.1 Australian trends

For Australia, mortality rates for cancer of the respiratory system[1] were low for both sexes in the earlier part of the twentieth century (Holman and Armstrong, 1982). The characteristic rise in rates began in the 1930s (see Figure 1.1)[2]. Males followed an exponential rise through to the 1960s, then rose more gradually to peak in the early 1980s and then declined in a near symmetric manner. For females, the rate of increase was considerably later and slower, with a suggestion that rates may have reached a plateau in the past five years.

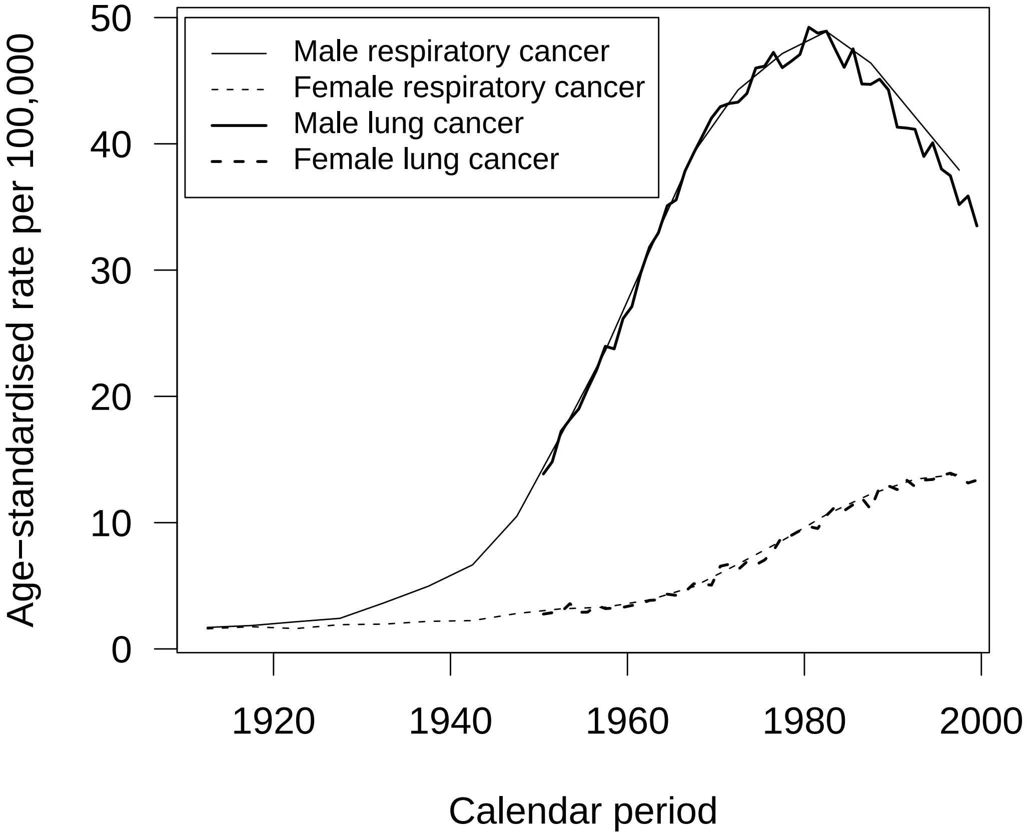


Figure 1.1: Mortality rates for lung and respiratory cancer in Australia, by sex, 1930–1999 (age-standardised to Segi's World population) (Data source: Australian Bureau of Statistics)

There was a close correspondence in mortality between lung cancer and cancer of the respiratory system. The main deviation from this relationship was for more

[1] *Cancer of the respiratory system* typically includes cancer of the lung, bronchus, trachea, pleura and mediastinum. *Lung cancer* typically includes only cancer of the lung, bronchus and trachea. See Chapter 2 for more detail.

[2] Respiratory cancer mortality rates are reported by five year periods and plotted at the period mid-point.

recent years. As detailed in Chapter 2, this deviation may be attributed to rising rates of cancer of the pleura as a result of asbestos exposure.

1.1.2.2 New Zealand lung cancer

The pattern for lung cancer in New Zealand is similar to that in Australia (Figure 1.2). As discussed earlier, incidence and mortality tend to follow a similar pattern because of poor survival.

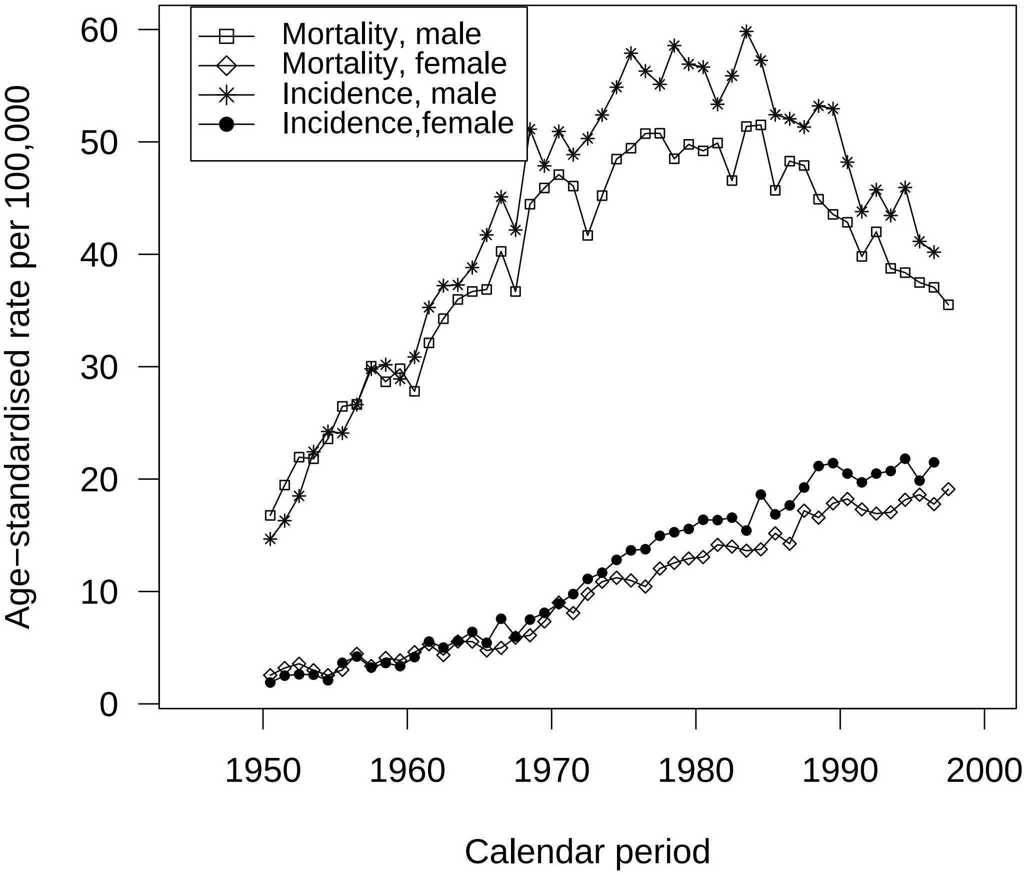


Figure 1.2: Lung cancer mortality and incidence in New Zealand, by sex, 1950–1998 (Data source: New Zealand Health Information Service)

The New Zealand cancer registration data were probably under-reported in the early years (Jim Fraser, personal communication). As an explanation, the system was voluntary and relied upon clinic-based registrations. Registrations began in private hospitals in 1972 and registrations based solely on death certification began in 1974.

1.1.2.3 International trends

There is considerable variation in lung cancer mortality rates between countries and over time (Miller, 1999). In the following section, age-standardised lung cancer mortality rates are presented for selected countries.

For males in the selected countries, rates rose through the third quarter of the twentieth century, reached a plateau during the 1970s–1990s, and declined in the latter part of the twentieth century (Figure 1.3). There is close agreement in rates over time for males in Australia and New Zealand. For British males, the mortality rates rose to higher levels than other countries and then began to decline earlier, dropping below the rates for US males in 1991.

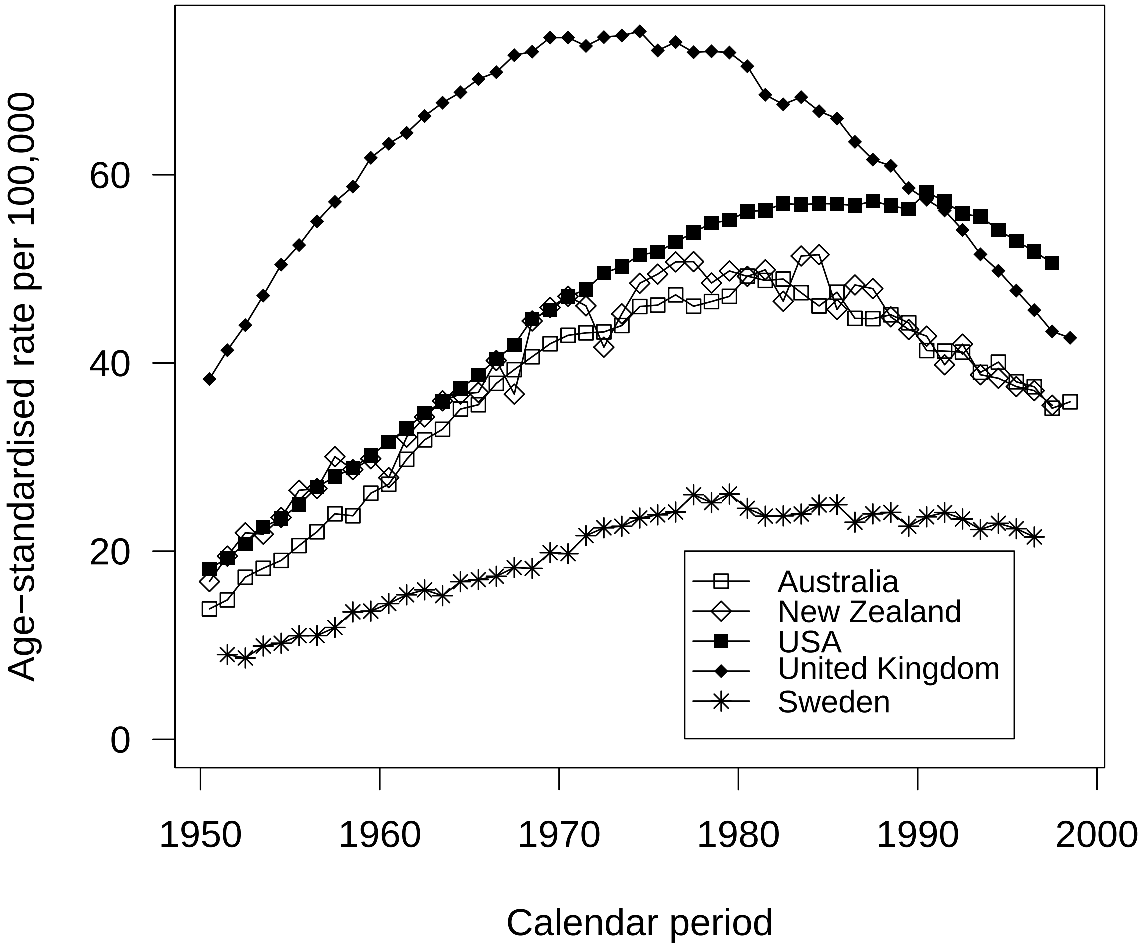


Figure 1.3: International comparison of lung cancer mortality rates for selected countries, males, 1950–1998 (age-standardised to Segi's World population) (Data source: WHO Mortality Database)

For females, the pattern for lung cancer mortality is different from that for males (Figure 1.4). As seen from the age-standardised rates for respiratory cancer (Figure

1.1), rates were low in the middle part of the twentieth century and then rose across the latter half of the century. British females again had higher rates earlier and then reached a plateau of gradual decline at around 1990. There is limited evidence for any decline in the age-standardised rates for females from other countries.

Rates among New Zealand females have tended to be higher than those for Australian females. This can in part be explained by higher levels of both smoking and lung cancer mortality in Māori women than in non-Māori women (Ministry of Health, 2000).

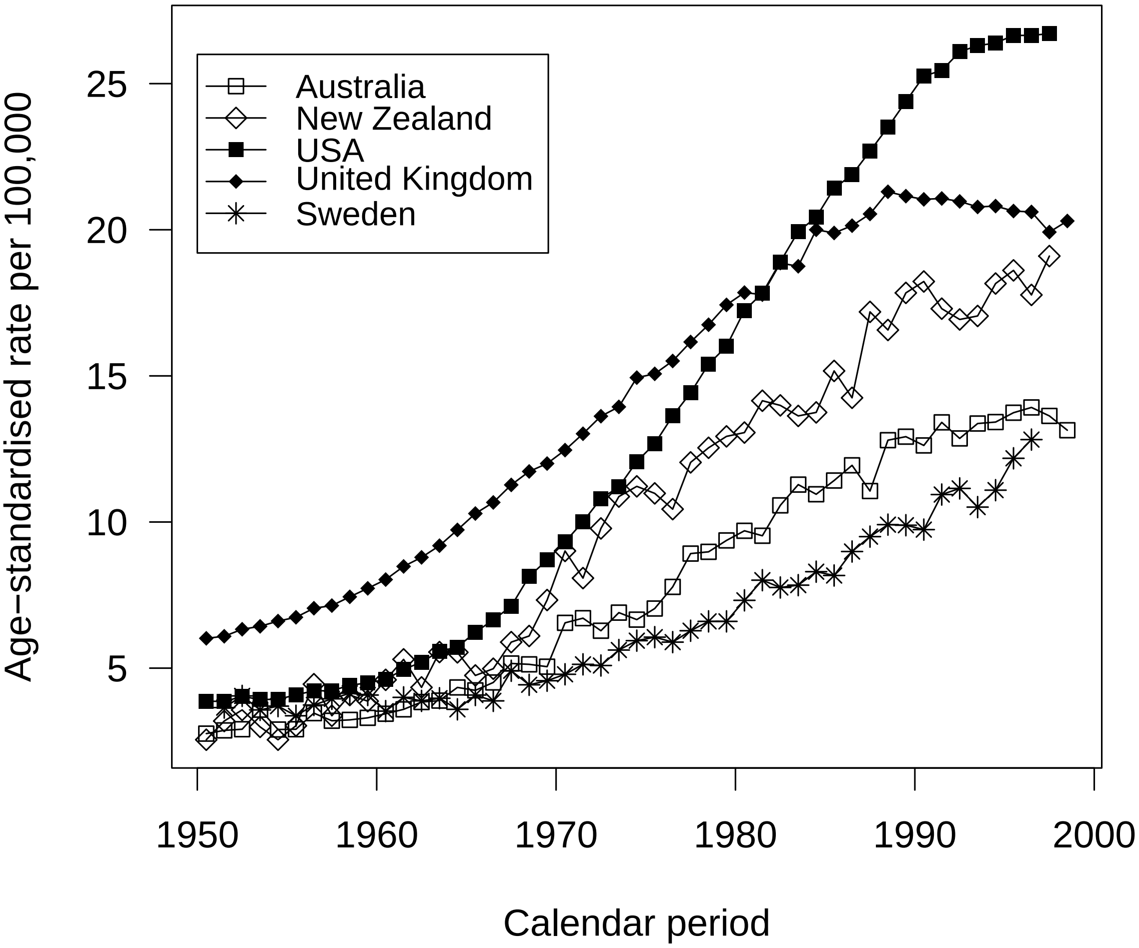


Figure 1.4: International comparison of lung cancer mortality rates for selected countries, females, 1950–1998 (age-standardised to Segi's World population) (Data source: WHO Mortality Database)

1.1.2.4 Age-specific lung cancer rates

Age-specific lung cancer rates can be represented using contour charts with coordinates for age and period (the Lexis diagram, historically reviewed by Keiding

(1990)). Rates for Australian males and females from 1930–1999 are shown in Figures 1.5 and 1.6, respectively. These results update an earlier presentation by Jolley and Giles (1992).

Before interpreting the charts, it is useful to revise their interpretation (Jolley and Giles, 1992; Robertson and Boyle, 1998). The contour lines represent lines of constant rate. The contour lines were estimated by smoothing the age-specific rates for five year age groups by five year cohorts (quinary quinquenia) using local likelihood regression (see Section 3.6.7). The smoothing approach is different from the spline interpolation suggested by Robertson and Boyle (1998) where the smoothing simplifies interpretation but sacrifices some detail. The smoothed age-specific rates for a given period, age or cohort are represented by slicing on a vertical, horizontal or (dotted) diagonal line, respectively.

For males, age-specific rates for younger ages (40–59 years) peaked in the early 1970s (1920–1929 cohort): this can be seen by tracking along a vertical line for a given age and identifying where a contour line is furthest to the left (in this case vertical). Rates at older ages (80–89 years) peaked in the late 1980s and early 1990s (1900–1909 cohort), which can be recognised from the vertical lines and the "hill top". There is evidence for both period and cohort effects, where the oldest cohorts had lowest rates and all age-specific rates increased during 1950–1969. The general outlook is for a decline in rates.

In contrast, lung cancer mortality rates for older females have not peaked (see Figure 1.6). Rates for younger females (40-49 years) became more stable during the early 1990s (1940–1949 cohort) some 20 years after the younger males stabilised. During the year 2000, the age-specific rates for females and males at younger ages were similar to within an order of magnitude, however the male rates rose considerably more quickly by age.

These changes in rates suggest that smoking exposure has varied appreciably by sex, age, period and cohort.

As a review of this section, the striking and well-recognised rise of lung cancer through the twentieth century has been presented, with an emphasis on results from Australia.

1.1.3 Tobacco exposure

Tobacco has been used in Australia since the arrival of Europeans in the eighteenth century (Tyrrell, 1999). Nicotine, the psychoactive ingredient of tobacco, produces effects including relaxation and alertness, which are probably the basis for its ad-

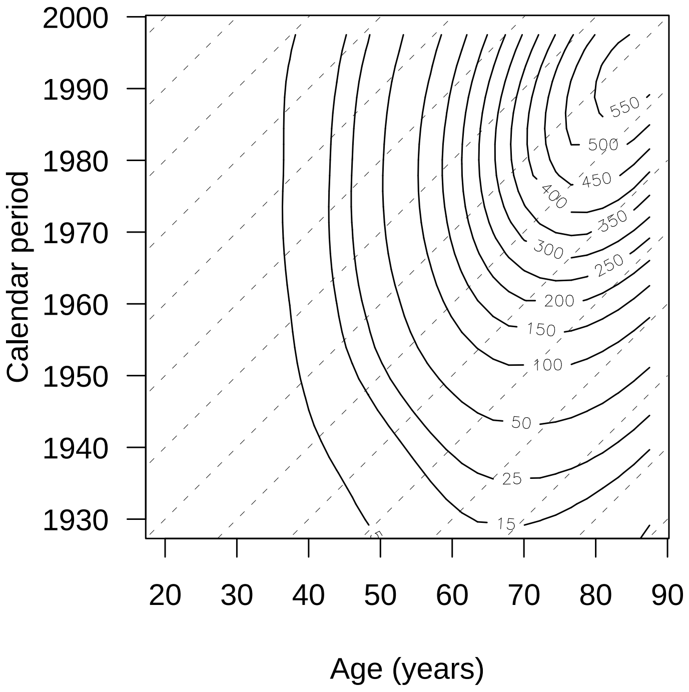


Figure 1.5: Cancer of the respiratory system: age-specific mortality rates per 100,000 among Australian males smoothed over 1930–1999 (Data source: Australian Bureau of Statistics)

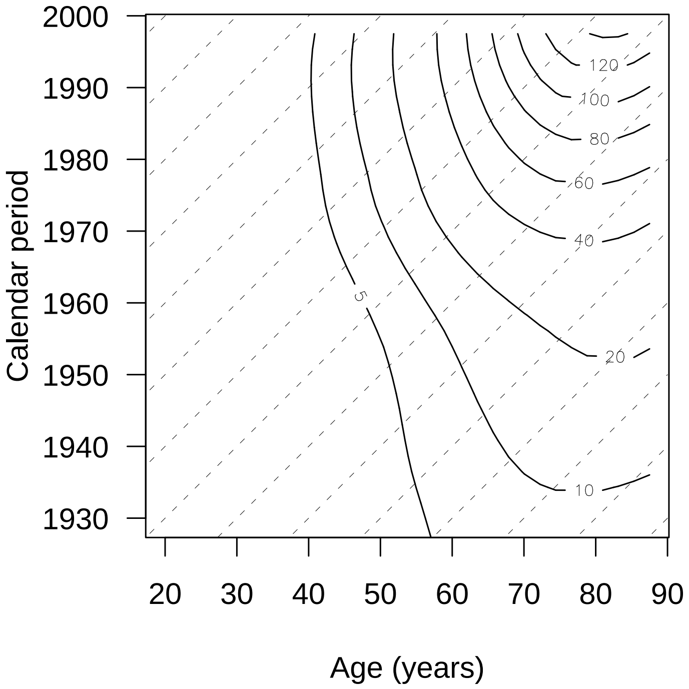


Figure 1.6: Cancer of the respiratory system: age-specific mortality rates per 100,000 among Australian females smoothed over 1930–1999 (Data source: Australian Bureau of Statistics)

dictiveness (Winstanley et al., 1995).

Population tobacco exposure is commonly described by total consumption and smoking prevalence. Consumption is often expressed as the quantity of tobacco products available for consumption per adult. Prevalence is expressed as the proportion of adults who are current smokers or the proportion of ever smokers who are former smokers. Adults are usually defined as the population aged over 15, 16 or 18 years.

An important premise for later analysis is that Australia and New Zealand have followed similar broad patterns of smoking exposure. The premise is motivated by the existence of novel data for New Zealand, including data from the Census, which will provide important parameters for analysis of the Australian data. New Zealand and Australia have similar histories, having both formerly been British colonies. The two countries are close both geographically and economically with open migration between them.

1.1.3.1 Tobacco consumption

Statistics for tobacco products available for consumption for Australia and New Zealand are collected from customs and excise. These data do not describe the demographics of the consumers. Tobacco consumption in Australia was relatively stable from 1910 to the Depression in the 1930s (see Figure 1.7).

Figure 1.8 illustrates the rapid changes in consumption by type of tobacco product in Australian and New Zealand over the twentieth century. In the earlier part of the twentieth century, a greater part of tobacco consumption was loose tobacco, smoked using either a pipe or using roll-your-own cigarettes.

There is strong agreement in the pattern of consumption of loose tobacco and manufactured cigarettes between Australia and New Zealand (Figure 1.8). This is somewhat surprising given the typically large differences in tobacco consumption experienced between developed countries (Todd, 1978) and over time (Thun et al., 1997)

Additional details on the consumption data are available in Section 2.2.

Tobacco consumption provides a long objective time series. The main potential source of bias is smuggling, which is difficult to estimate accurately. Moreover, any attempt to divide tobacco consumption across ages and between sexes is limited by the available smoking prevalence data.

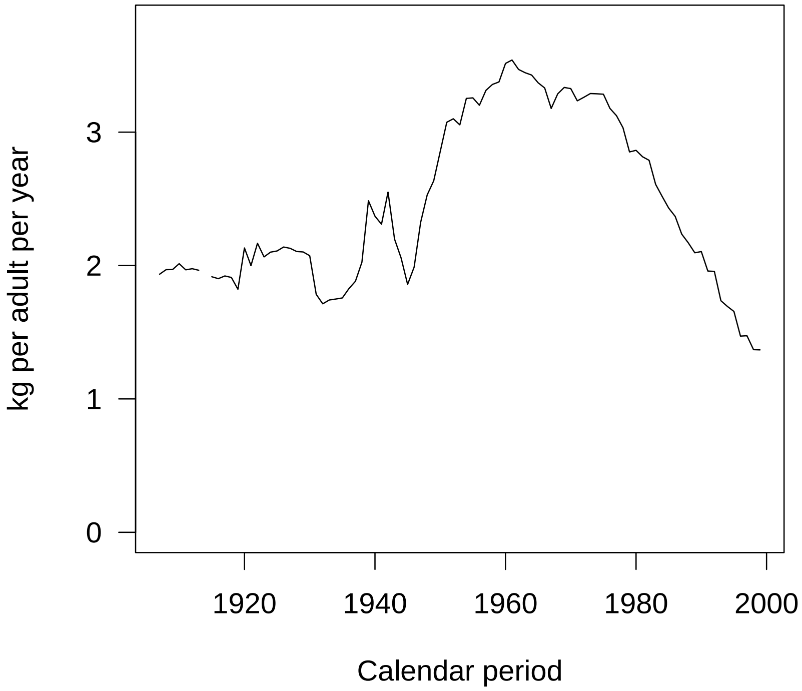


Figure 1.7: Tobacco products available for consumption in Australia, 1907–1999 (Data source: Australian Bureau of Statistics)

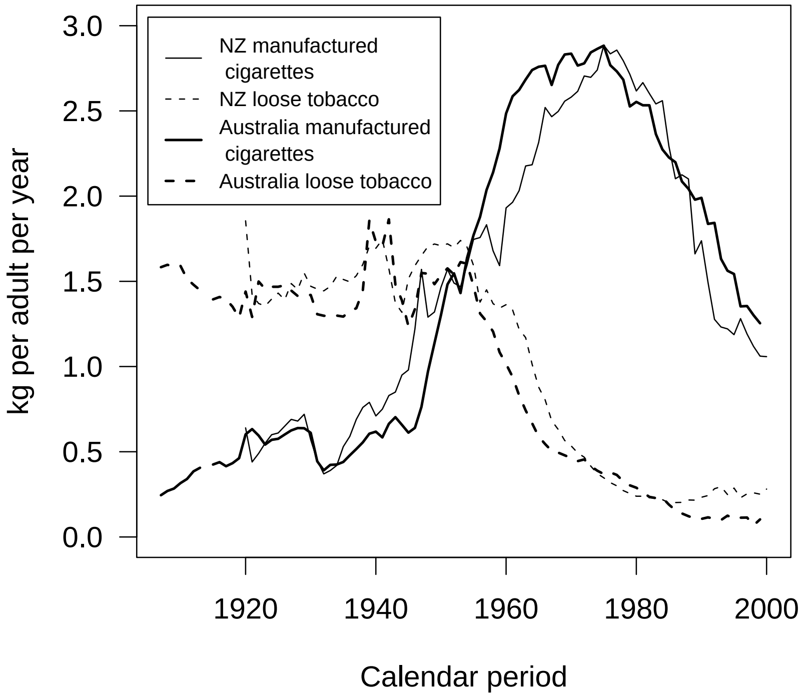


Figure 1.8: Manufactured cigarettes and loose tobacco available for consumption for Australia and New Zealand (Customs and Excise) (Data source: Australian Bureau of Statistics and Statistics New Zealand)

1.1.3.2 Smoking behaviour

The most common sources for data on smoking behaviour are surveys that ask respondents to self-report their smoking status. Few population-based smoking data based on biomarkers are available.

Estimated consumption based on self-report tends to under-estimate consumption relative to consumption based on customs and excise data. The consistency of under-reporting has been questioned (Warner, 1978), however more recent research suggests that self-reported smoking from population-based samples may have a consistent bias over time and therefore represents a valid measure of smoking behaviour (Patrick et al., 1994).

Smoking is a complex behaviour with a variety of contexts including that of the individual, of society, of economics, of addiction and of health (Giovino et al., 1995). Any description of the individual's history of smoking may also be complicated. A person may begin experimenting with smoking at a certain age, then possibly begin smoking on a regular basis, and then possibly stop smoking. The pattern of smoking may vary considerably, including frequency of smoking, the choice of products smoked, and inhalation patterns. Moreover, smoking cessation is poorly defined, because a former smoker may take up smoking again after a period of cessation and make several more quit attempts.

Definitions For ease of presentation and analysis, this complexity is often summarised. At a population level, it is common to categorise smoking status into current, former and never smokers, which requires a definition for smoking (Mattson and Kessler, 1987).

For frequency of smoking, the definition tends to include whether the respondent smokes a particular tobacco product or smokes the tobacco product "regularly" or "daily". Former smokers are those who have "ever" or "used to" smoke with the given frequency but do not at present, possibly with the condition that they have smoked 100 or more cigarettes or equivalent in their lifetime.

For a definition of smoking, products typically include either all tobacco products or just cigarettes, both roll-your-own and manufactured. There is often a distinction between cigarette smokers and smokers of other tobacco products as there is evidence suggesting cigarettes smokers tend to be at a higher risk for some diseases (Doll and Hill, 1966).

Australian smoking prevalence The Anti-Cancer Council of Victoria (ACCV) collated data on Australian smoking prevalence for the period 1974–1995. The definition of smoking used by the ACCV is “regular” smoking of any tobacco products.

The age-standardised prevalence of current smoking for adults aged 16 years and over is presented in a similar manner to rates for lung cancer. Male smoking prevalence declined across the period, although the downward trend appears to have abated at the end of the period (Figure 1.9). For females, smoking prevalence varied less over time, with similarly limited evidence for a plateau at the end of the period.

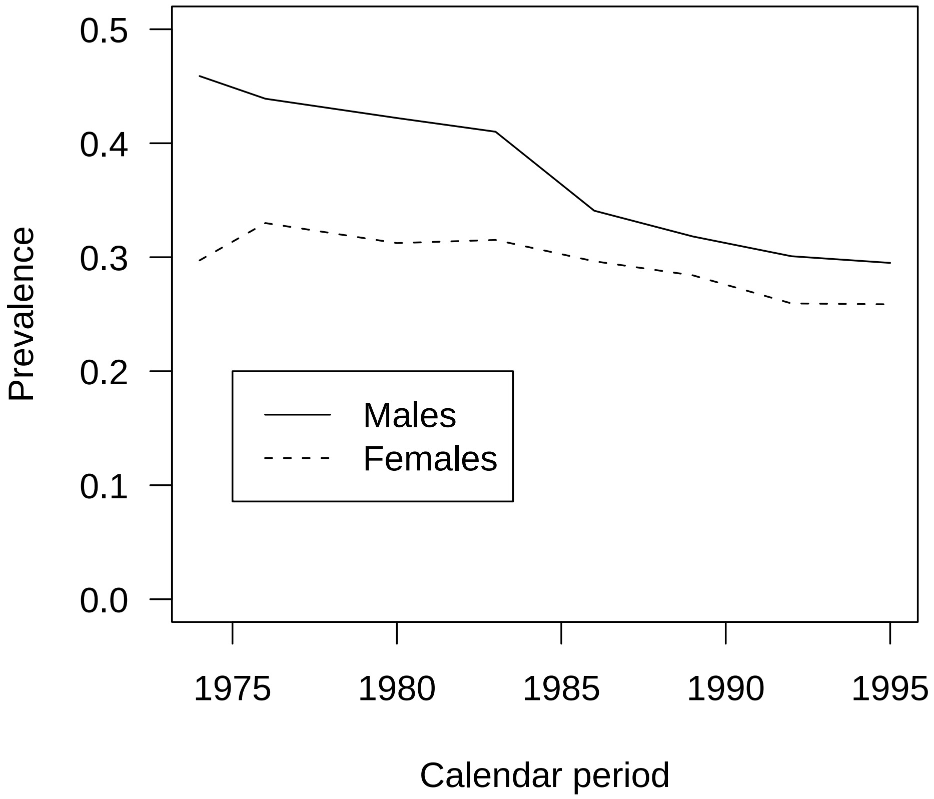


Figure 1.9: Age-standardised prevalence of current smoking in Australian adults by sex, 1974–1995 (age-standardised to Australia 1991) (Data source: Anti-Cancer Council of Victoria)

The pattern of change varied by age and sex. Males have shown a consistent decline in current smoking prevalence across the different ages over time (Figure 1.10). In particular, prevalence dropped substantially for older males.

In contrast, females showed limited reductions in current smoking prevalence over the period (Figure 1.11). In particular, prevalence at younger ages changed little over the period. Prevalence for the older females stayed low, as seen at the beginning of the period.

What proportion of ever smokers have quit? This can be measured by the so-called *quit-ratio* or *prevalence of cessation*, which is the proportion of “ever smokers” who report as having quit smoking. This measure is age-dependent, where the preva-

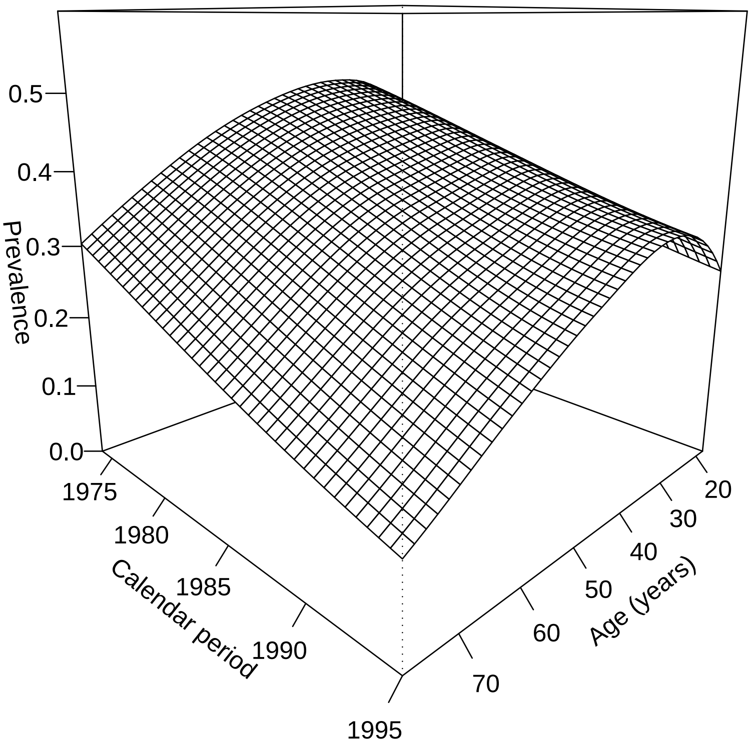


Figure 1.10: Prevalence of current smoking by age and period, Australian males during 1974–1995 (Data source: Anti-Cancer Council of Victoria)

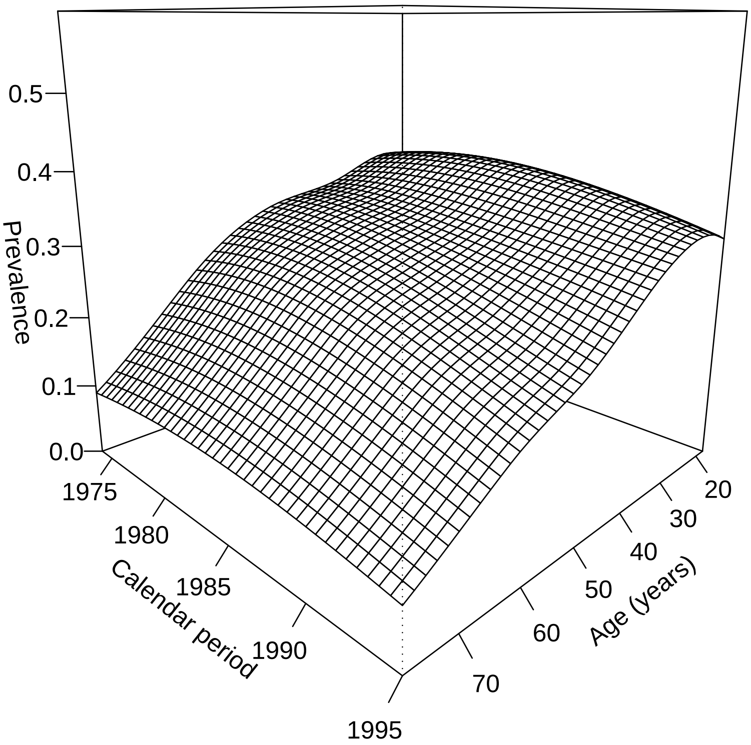


Figure 1.11: Prevalence of current smoking by age and period, Australian females during 1974–1995 (Data source: Anti-Cancer Council of Victoria)

lence of cessation is expected to rise with age due to increased opportunity to quit and due to differential mortality, because former smokers are more likely to survive than current smokers. For a methodological discussion, see Section 3.6.8. Across the period 1974–1995, the prevalence of cessation for males tended to increase for a given age, suggesting increasing rates of cessation (Figure 1.12).

The pattern for females is slightly more complex (Figure 1.13). The prevalence of cessation for women aged around 50 years in the middle to late 1970s (1920–1925 birth cohort) was appreciably lower than for younger or older age groups. However by 1995 there was less evidence for this "basin".

In summary, there was wide variation in smoking behaviour in Australia during the twentieth century. Such variation has been observed elsewhere, including Europe (Graham, 1996) and the United States of America (Thun et al., 1997). The two large US cohort studies by the American Cancer Society also observed changes in daily cigarette consumption, smoking duration and machine-measured tar content (Thun et al., 1997).

1.1.4 Lung cancer aetiology

Smoking is a well-established causal factor for lung cancer. Surprisingly, several aspects of the aetiology are contentious. The following is a brief review of relevant relationships for population lung cancer mortality modelling with some indication of the points of contention. For more detailed reviews, see Samet (1994) and Peto (1977). Chapter 10 also provides some discussion in the context of multi-stage models for carcinogenesis and lung cancer rate modelling.

Determinants of lung cancer include active smoking of tobacco and marijuana, environmental tobacco smoke, air pollution, asbestos, ionising radiation, occupational exposures, and the protective effects of nutrition and diet (Samet, 1994). There are also important interactions between tobacco smoking and both asbestos and ionising radiation, with interactions that are additive to multiplicative (Saracci, 1987). Asbestos exposure increases the risk of cancer of the pleura (Xu et al., 1985), suggesting that cancer of the pleura be excluded from subsequent lung cancer models (see Chapter 2). Natural exposure to radon in most Australian homes is very low and is expected to have little impact on lung cancer risk (Langroo et al., 1991). Limited Australian data are available for changes in risk over time due to the other factors.

We focus now on the relationship between tobacco smoking and lung cancer.

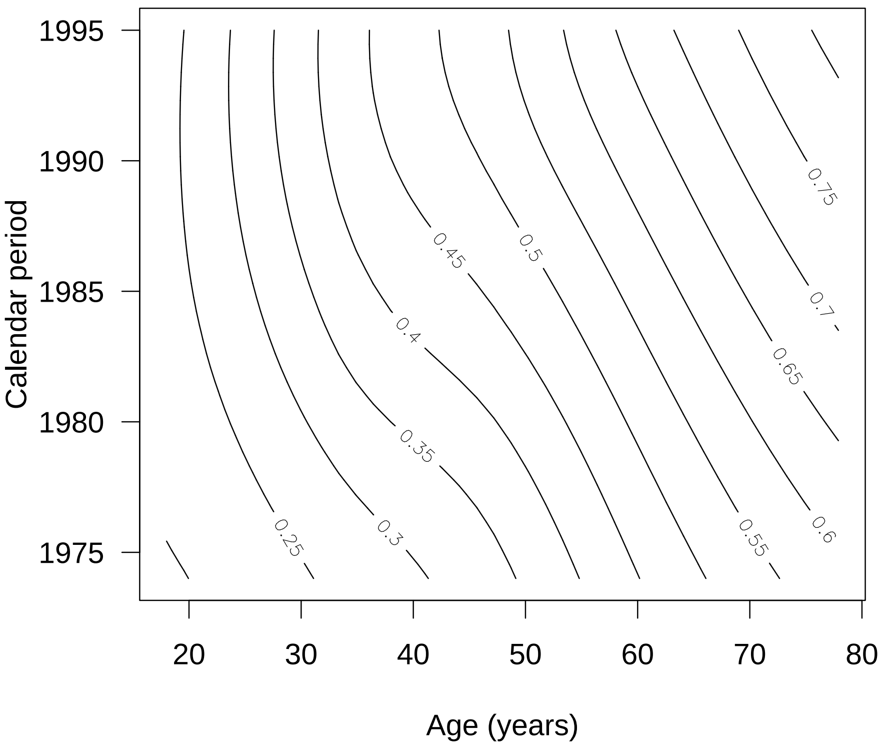


Figure 1.12: Prevalence of cessation by age and period, Australian males during 1974–1995 (Data source: Anti-Cancer Council of Victoria)

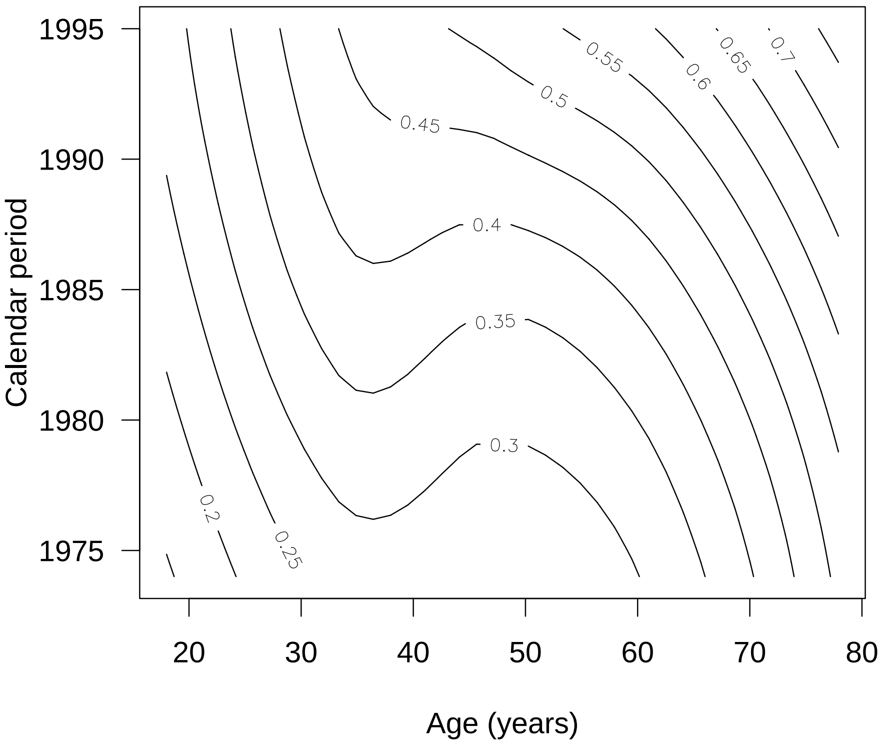


Figure 1.13: Prevalence of cessation by age and period, Australian females during 1974–1995 (Data source: Anti-Cancer Council of Victoria)

1.1.4.1 Rate ratios

A common assumption in epidemiology is that rate ratios are constant between sexes, over ages and over time. This assumption is made in the absence of any evidence for risk interactions between the different factors. Data to address possible risk interactions are available from two large cohort studies supported by the American Cancer Society (Thun et al., 1997). These studies, called the Cancer Prevention Study (CPS) I and II, were started in 1966 and 1986.

Rate ratios for current to never smokers for the two cohort studies by sex are shown in Figure 1.14. The smoothed trends in rate ratios have been interpolated and smoothed using local likelihood estimation on the rates[3], with inference using generalised linear models[4]. A log scale has been used because the rate ratios are in general appreciably greater than one. There is consistent evidence for the rate ratios being concave with age and that the CPS-II rate ratios reached a peak at younger ages than the rates from CPS-I.

Male rate ratios increased by a factor of 1.73 (95% confidence interval: 1.30–2.31) between CPS-I and CPS-II. The increase in female rate ratios between CPS-I and CPS-II (3.61, 95% confidence interval 2.82–4.61) was greater than for males ($p < 0.001$).

Comparing CPS-I and CPS-II, both male and female smokers tended to smoke more cigarettes per day (Thun et al., 1997). Moreover, female smokers from CPS-II tended to consume more cigarettes per day and reported inhaling the tobacco smoke more deeply. These factors may explain the greater part as to why female lung cancer rate ratios increased at all ages between the two studies.

In summary, the lung cancer mortality rate ratios for current smokers compared with never smokers of the same age were higher in the later study. Not only had more people smoked, but the changing pattern of smoking changed the level of risk. This strongly suggests the need to adjust for cumulative smoking exposure, which was observed to change between CPS-I and CPS-II.

1.1.4.2 Duration

Aetiological studies show a strong relationship between duration of smoking and risk of lung cancer (Doll and Peto, 1978). This relationship remains after adjustment for

[3] For the local likelihood regression, age and an interaction between age and an indicator for smoking were included as covariates.

[4] For the generalised linear models, main effects included indicators for study and sex (where included) and quadratic terms for age. There were also interactions between the main effects and an indicator for smoking.

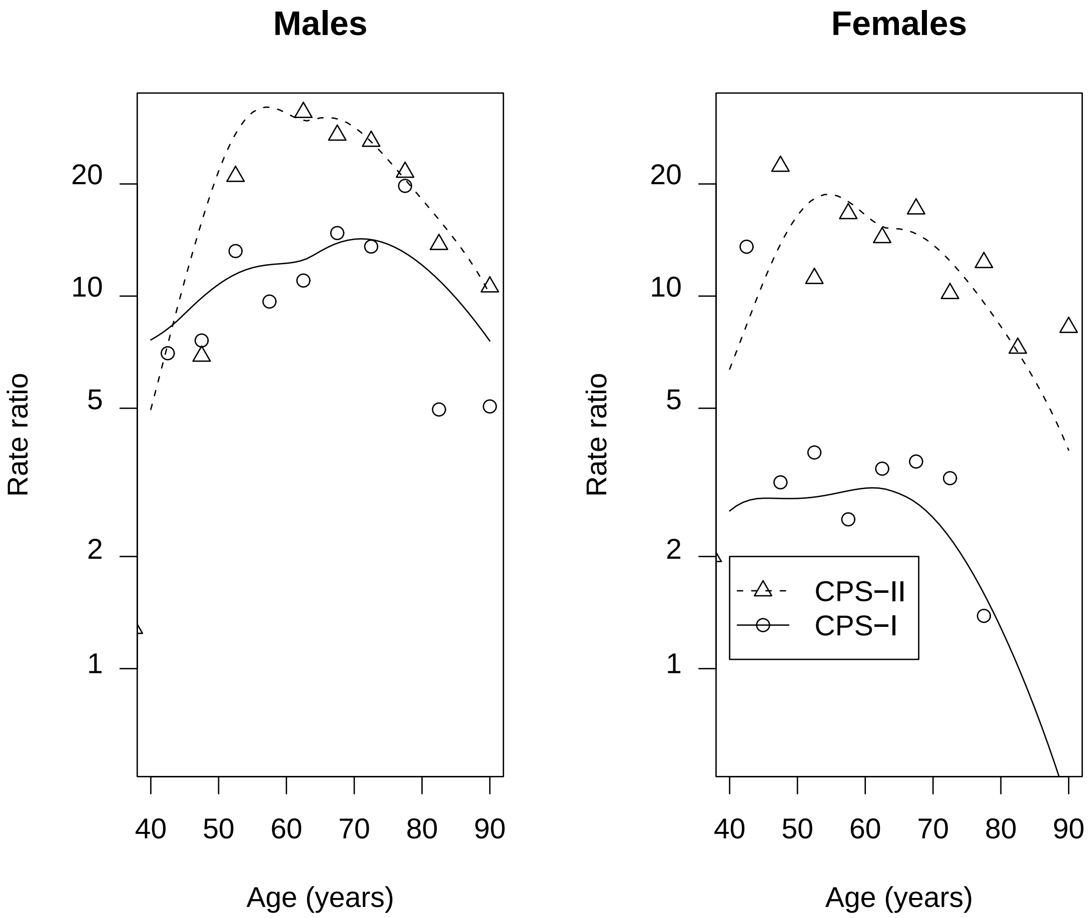


Figure 1.14: Observed and smoothed age-specific lung cancer mortality rate ratios for current smoking versus never smoking from Cancer Prevention Studies (CPS) I (1966) and II (1986) by sex (Data source: Thun et al., 1997)

average dose.

An important observation by Doll (1971) was that smoking duration for current smokers was related to lung cancer incidence rate by the fourth or fifth power, which was a similar power relationship between age and lung cancer incidence for never smokers. This association was an early motivation for the application of multi-stage models of carcinogenesis to lung cancer modelling (see Chapter 10).

As discussed in the previous section, females from CPS-II started smoking earlier and had a longer average duration of smoking compared with their counterparts in CPS-I (Thun et al., 1997). There was limited change in estimates of male smoking duration between CPS-I and CPS-II.

1.1.4.3 Age

The difficulty in differentiating between the effects due to age and those due to duration is well described in the monograph by the International Agency for Research on Cancer (1986): "The effects of the duration of smoking are so strong, and so closely correlated with age, that it is virtually impossible to determine exactly whether ageing *per se* has any independent effect on excess lung cancer rates among people of different ages who have all smoked similarly for a similar number of years. If age has any independent effect, however, this would be small compared with the accumulative effect of duration of smoking."

1.1.4.4 Dose

There is a linear to quadratic relationship between the number of cigarettes smoked per day and lung cancer incidence (Doll and Peto, 1978). As will be discussed further in Chapter 10, empirical evidence supports a linear relationship, which is at contradiction with the multi-stage theory (Armitage, 1971). In an unpublished review document, Lee (1995) suggests reasons for observing a linear rather than the expected quadratic effect, including imprecise reporting, risk heterogeneity in the population, and a possible high risk threshold.

There is also some debate as to whether early or late dose has more influence on lung cancer. Brown and Chu (1987) found evidence for a stronger late stage effect on a multi-stage model for smoking and lung cancer. Lee (1979) suggested that, given the available data, it was equivocal as to whether early or late dose had a greater effect on lung cancer risk.

Doll and Peto (1978) made the interesting observation that the heaviest smokers from the British Doctor's Study were found to be at similar risk to moderately heavy smokers. One consequence is that the dose-response curve may be non-linear.

1.1.4.5 Cumulative consumption

At the individual level, Thomas (1987) suggests that cumulative measures of tobacco consumption, such as pack-years, are a limited measure of lung cancer. Thomas argued that the average risk for a person with the same cumulative consumption was higher if they had smoked for a longer duration at a lower dose. However Whittemore (1988) found that in practice the pack-years function performed quite well.

At the population level, cumulative consumption has proved to be a useful aggregate measure of tobacco exposure (Stevens and Moolgavkar, 1984; Yamaguchi et al., 2000). One possible explanation is that average dose and duration for a population may change slowly over time, so that cumulative consumption can be used effectively to represent lung cancer risk for both smokers and for the general population.

1.1.4.6 Filters and tar

There is consistent evidence that the use of cigarette filters is associated with a reduction in the risk of lung cancer (Samet, 1994). The evidence is more equivocal for cigarettes that are measured as having lower tar. Several investigators have found that the association with tar is attenuated after adjustment for the quantity smoked (Lubin et al., 1984). One explanation for this effect is that smokers of cigarettes with lower tar and nicotine levels tend to compensate by smoking more cigarettes or inhaling more deeply (Wilcox et al., 1988; Burns et al., 2001).

There has been a recent debate about testing of tar and nicotine levels, because changes in cigarette construction may give inaccurate yield estimates (Wilkenfeld et al., 2000; Burns et al., 2001). Consequently, estimates of tar exposure over time may be invalid.

1.1.4.7 Inhalation

Depth of inhalation is closely related to the risk of lung cancer (Samet, 1994). Compared with pipe and cigar smokers, cigarette smokers tend to be at increased risk of lung cancer because the smoker tends to inhale more deeply. Moreover, with the increasing use of low tar cigarettes, smokers may be inhaling more deeply to increase nicotine yield (see previous section).

1.1.4.8 Gender

An important aspect of the lung cancer epidemic has been the different trends between males and females. Although the gross changes in the epidemic can largely be ascribed to differential exposure between males and females, there is some debate as to whether females are more susceptible to lung cancer for a given level of smoking exposure (Prescott et al., 1998; Marang-Van de Mheen et al., 2001). Models for smoking and lung cancer would ideally be performed separately for males or females, or would formally test for interactions.

1.1.4.9 Smoking cessation

The risk of lung cancer declines following cessation of the habit (U.S. Department of Health and 1990). However there has been some debate over the form of the change for the British Doctor's Study, which has been summarised by Moolgavkar et al. (1989):

"In fact, interpretation of the data on risk among exsmokers is quite controversial, and the only certainty is that this risk lies somewhere in between the risk among continuing smokers and that among nonsmokers."

A useful approximation for the risk for former smokers is the sum of the cumulative risk for the time as a current smoker and the background age-related risk (Peto, 1977).

One complication of this approximation is that those who stop smoking may be different from the average smoker, as in smoking cessation in response to illness. As a consequence, mortality rates for recent quitters may be elevated compared with those for current smokers. This heterogeneity makes prediction of the lung cancer risk for former smokers difficult.

1.2 Methodological issues

There is good evidence that tobacco control interventions improve population health (Public Health Commission, 1994; Commonwealth Department of Health and Family Services 1998). However the public health purse is limited. Methods are required to predict how specific interventions will affect health and what the future holds regardless of intervention. This has importance for determining health priorities and for health planning.

1.2.1 Predicting the impact of interventions

In comparing interventions for determining health priorities, the effect of a specific intervention on health needs to be considered against the associated costs. By taking a risk factor approach, the effect of an intervention on health can be further decomposed into how an intervention affects a risk factor, and how changes in the risk factor affect health.

Health economics and the associated cost assessment will not be considered further in this thesis. Questions of health promotion and health policy, assessing how different interventions change risk factor distributions, will also not be considered further.

As a justification for the risk factor approach, lung cancer is a rare outcome and measurement of pre-clinical states using population screening is not presently an option. The most common proxy outcome for lung cancer risk is then a risk factor itself: smoking tobacco. Therefore in this case, the risk factor approach is the most practical approach to address the potential impact of an intervention on lung cancer.

The choice of disease and risk factor fits well within the burden of disease framework. This is arguably because lung cancer is a common cancer and a large proportion of lung cancer is *potentially* preventable by a reduction in population smoking. However there has been some debate on the role that burden of disease and epidemiology have played in health priority setting, where resources have been taken away from other efforts (Murray et al., 1994; Mooney et al., 1997).

1.2.1.1 Attribution

The link between burden of disease and risk factors has to date lent heavily on the use of population attributable fractions and their derivatives (Murray and Lopez, 1997a). One approach estimating the potential impact of an intervention is to use a population attributable fraction as an elasticity, where a small change in exposure will be represented by a scaled change in burden of disease (Hill, 1996). The associated counter-factual question is "had there been a proportional drop in exposure, what would have been the decrease in disease rate?"

An important assumption is that changes from a current intervention will be expressed in a change in disease in the near future. This assumption will generally not hold for smoking and lung cancer.

1.2.1.2 Potential impact fraction

A closely related concept is the *potential impact fraction*, which is the proportional change in a rate given a different exposure distribution (Morgenstern and Bursic, 1982).

However for cancers that are causally associated with tobacco smoking, population risk is associated with measures of cumulative smoking exposure. Moreover, there have been changes in the rate ratios, so population attributable fractions based around predicted changes in smoking prevalence and fixed estimates of the rate ratios will not accurately represent the actual change.

Gunning-Schepers and Barendregt (1992) discuss the importance of time in predicting the effect of interventions. The authors propose a modification that linearly scales the potential impact fraction over time and includes estimates of any secular trend in the rate ratios. A software implementation of this approach is available (`Prevent`). This approach incorporates smoking initiation and cessation, however the model has little capacity for including smoking exposure more than 15 years in the past.

This modelling approach is a significant step forward, but there are several concerns. First, the linearity and choice of interval for the change in the potential impact fraction are open to question. Second, the trend in the rate ratios is dependent on cumulative smoking exposure. This requires good estimates of smoking exposure and modelling for the association between cumulative exposure and lung cancer. This thesis will investigate methods to improve the quality of available estimates.

1.2.2 Prediction regardless of intervention

Another issue, separate from the setting of health priorities, is the planning of health services, which is reliant upon predictions of the number of new lung cancer cases. A variety of approaches have been used to predict lung cancer incidence and mortality. Most of these approaches do not explicitly include information on smoking exposure. A brief review of these models, with some consideration of implementation factors, is now given.

1.2.2.1 Age-period-cohort models

A common approach to lung cancer prediction is to use Poisson regression with the lung cancer rates, taking account of age, period and cohort effects. Age-period or age-cohort models can be fitted using generalised linear models when the covariates are polynomials or factors.

Considerable attention has been given to age-period-cohort models (Clayton and Schifflers, 1987), although their use has been criticised (Moolgavkar et al., 1998). The main difficulty is that age, period and cohort are linearly related, so that a constraint is required to find estimable parameters. The models require careful specification of the predicted effects for age, period and cohort, which can be useful for modelling counterfactual or future scenarios (e.g. La Vecchia et al., 1988; Taylor et al., 2001).

Lung cancer projections using these models have been used in Australia (Taylor and McNeil, 1997), New Zealand (Cox, 1995) and elsewhere (e.g. Coleman et al., 1993; Nam et al., 1996; Jee et al., 1998).

1.2.2.2 Simple non-linear models

Scandinavian cancer incidence predictions have been based on a linear period shift of age-specific rates (Dyba et al., 1997; Dyba and Hakulinen, 2000). This non-linear model is mathematically simple and has the useful property that the age-period and age-cohort models are mathematically equivalent. Another useful property of these predictions is that the confidence intervals are small. However fitting the models requires programming for iteratively re-weighted least squares or generalised non-linear models (Lindsey, 2001).

1.2.2.3 Modern regression

By using a geometric argument, age, period and cohort effects cannot effectively be disentangled, as there are three parameters being used to describe a surface defined on two dimensions. Parametric age-period and age-cohort models acknowledge this relationship, however they suffer from assuming that changes are independent. A more attractive approach is to model the surface using modern regression methods.

Several tool-kits exist for Poisson distributions, including generalised additive models (Hastie and Tibshirani, 1990) and local likelihood estimation (Loader, 1999). The approach is conceptually simple: fit a non-parametric surface and extend it out in time. The main technical difficulties are model building and inference.

This approach using local likelihood fit has been employed in Figures 1.5 and 1.6. This approach is used to predict the age-specific lung cancer mortality rates, summarised using age-standardised rates in Figure 1.15. Two models were used, using either age and period or age and cohort as covariates.

This model predicts that the male lung cancer rates will continue to decline, while the female lung cancer rates may begin to slowly decline during the calendar period 2000–2010. If the models are representative, then male rates may approach female rates around 2015. The age-cohort and age-period models gave consistent results for males. However for females the two model fits provided qualitatively different results, where it is not clear how quickly female rates will decline. A second issue is that the level for male rates may approach those rates seen among females, suggesting that male rates may begin to plateau.

Similar estimates could also have been achieved using a two-dimensional spline in a generalised additive model.

Age-cohort models using generalised additive models with additive one-dimensional splines did not give satisfactory results for females, where the model did not describe well the observed changes from Figure 1.6. This has implications for using generalised linear models with only main effects, as such models would not well represent the complex changes in rates between ages and cohorts over time.

1.2.2.4 Bayesian methods

Bayesian approaches to age-period-cohort modelling are equivalent to smoothing over the parameters. Two approaches have recently been used in the literature: the

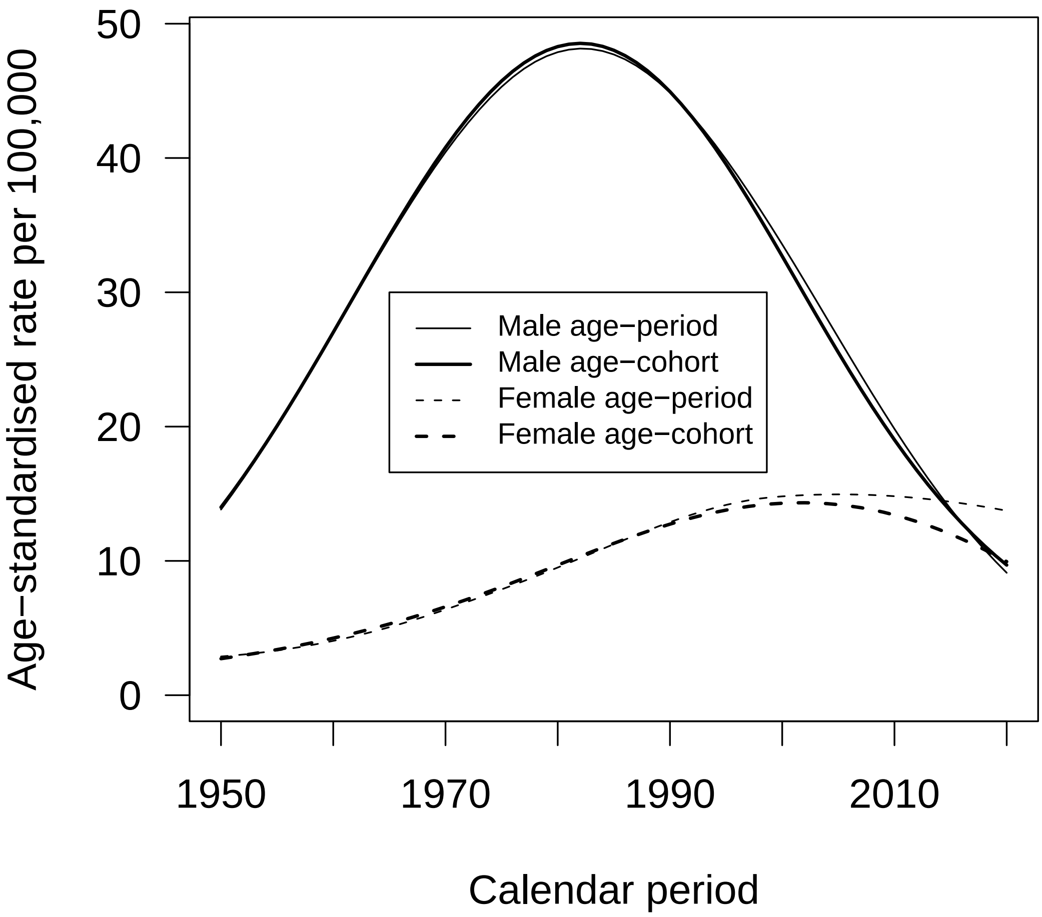


Figure 1.15: Fitted and projected lung cancer mortality rates using local likelihood regression models, by sex, Australia 1950–2020 (Data source: Australian Bureau of Statistics)

use of generalised linear mixed models with an autoregressive prior (Bray et al., 2000); and the use of a more complex model that requires specialised software (Knorr-Held and Rainer, 2001). Problems associated with mixing suggest that one or more other models should support the Bayesian solution.

An application of the method due to (Bray et al., 2000) using `BUGS` software suggests that the confidence intervals are very wide in comparison with simple estimators proposed in Section 1.2.2.2.

1.2.2.5 Inclusion of exposure information

Where the data are available, exposure data can improve cancer predictions (Moolgavkar et al., 1998). A good example is smoking and lung cancer, where a good aetiological model is available and reasonable exposure data are available (e.g. Townsend, 1978; Stevens and Moolgavkar, 1984; Doyle, 1985; Mantel et al., 1986; Tolley et al., 1991;

Swartz, 1992; Holford et al., 1996; Lee and Forey, 1998; Haldorsen and Grimsrud, 1999; Yamaguchi et al., 2000).

The complexity of these models can vary. The main difficulty with this approach is obtaining accurate exposure estimation. However any analytical efforts will be rewarded with providing a mechanism to also model for the effects of interventions.

1.2.3 A novel approach

Lung cancer models have been constrained by the validity and precision of the available exposure data. As a novel approach to improve the quality of the exposure data, it is proposed to develop and apply a multi-state smoking model. The approach is based on a cross-fertilisation of themes, including:

- Modern survival analysis and smoothing methods
- Non-linear mathematical models
- Multi-stage models of carcinogenesis.

This marriage of statistical estimation, mathematical modelling and epidemiology has been used effectively for modelling emerging infectious diseases, including HIV/AIDS (see Chapter 3 for a review). One limitation of this marriage is that the formulation of the mathematical model may be elaborate while the estimates are poorly validated, or the model may be poorly specified while the estimates are well validated.

Importantly, the multi-state model provides a systematic framework for estimation of a range of parameters including prevalence, duration and dose. Estimates for smoking prevalence, dose and duration are applied to lung cancer mortality to estimate parameters for the lung cancer risk equation.

This approach is closed related to that by Tolley et al. (1991), who used a multi-state model with states for never smokers, and for current and former smokers by smoking duration. This thesis advances this approach by developing more sophisticated rate models and methods for estimating a range of smoking parameters.

The approach lends itself to predicting outcomes of tobacco control interventions. The approach explicitly includes cessation and initiation rates and dose, allowing for intervention scenarios that affect one or more of these smoking parameters.

1.3 Thesis aims and outline

1.3.1 Aims

The general aims of the thesis are:

- To develop a multi-state model for population smoking behaviour
- To develop methods to estimate the parameters of the multi-state model under different sampling schemes
- To apply these methods to estimate smoking parameters
- To estimate the lung cancer risk function with parameters for smoking dose and duration
- To predict future lung cancer mortality.

The applications are focused on national estimates for Australia, with some results for New Zealand. The main disciplines are epidemiology and biostatistics.

1.3.2 Research strategy

The research strategy followed in this thesis included the development of a multi-state smoking model in order to estimate various smoking parameters. The model included states for never, current and former smokers and a state for death. Transitions were included for uptake, cessation, recall error by former smokers as never smokers and differential mortality.

The main analytical challenge was to validly estimate the smoking transition rates given data limitations. Rather than depend on broad assumptions, analysis was restricted to cohorts for whom good estimates were available. Estimation required a variety of data sources. For all cause mortality, this required a review of the literature and a method to estimate rate ratios that were specific to Australia and New Zealand over time.

Transition intensities for smoking initiation and cessation for cohorts before 1974 were estimated from retrospective data. For cessation rates since 1974, estimates were taken from cross-sectional data. New Zealand census data were used to estimate recall error rates and to validate retrospective cessation rate estimates.

These transition intensities were brought together for predicting backwards in time in order to adjust estimates for differential survival and for recall error. Then the model was forward projected in time to estimate various smoking exposure parameters.

The smoking parameters were used as inputs to lung cancer regression models. The fitted regression model were then applied to predicted smoking exposure parameters to predict future lung cancer mortality rates.

The lung cancer mortality models were fitted for ages 35–69 years and cohorts born from 1910.

1.3.3 Outline

Chapter 2 will describe the range of data sources that will be used throughout the thesis.

Chapter 3 introduces the multi-state smoking model and develops the relevant theory. Both the mathematical models and the estimation procedures are considered.

Chapter 4 provides a review of all cause mortality rate ratios, which are an important component of the multi-state model.

Chapter 5 provides retrospective estimates of smoking initiation and cessation rates using three different data sources. Chapter 6 provides results based on fitting a non-linear dynamic model to New Zealand Census data. Validation of retrospective estimates is undertaken, and estimates for rates of recall error by former smokers as never smokers are provided. Chapter 7 uses current status data from Australia to estimate cessation rates. Chapter 8 outlines how the multi-state model was completely specified for Australia and for New Zealand. Smoking parameters that are then derived from the model include differential survival and duration of smoking by current and by former smokers.

Chapter 9 describes age-specific consumption of smoking for Australia and New Zealand for recent years based on total consumption of tobacco divided between ages and sexes. Similar estimates for earlier years are proposed. Estimates for tar content are discussed.

Chapter 10 reviews the theory and previous applications for lung cancer risk functions dependent upon smoking exposure. Several lung cancer risk functions were fitted to the exposure data. Projections of exposure under different smoking scenarios were applied to the fitted lung cancer mortality rate equations to project for lung cancer mortality.

A summary of the results and a discussion of analytical issues and future applications are discussed in Chapter 11.

Two appendices have also been included. Appendix A is a report to the Commonwealth by the author and Associate Professor Richard Taylor on lung cancer projections using a simpler method. Appendix B is a short paper that applies methods from the thesis to show that US smoking prevalence may not continue to decline.

Chapter 2

Data sources

Abstract

Data are described for mortality and cancer incidence, populations, tobacco consumption and information on smoking behaviour from different population-based surveys. Questionnaires on smoking are included.

2.1 Mortality, cancer registrations and populations

Incidence data of interest include mortality data for all causes of death and for specific causes, and cancer registration data for lung cancer. Mortality and population data have been grouped together here because they tend to be collated by the same agency. Cancer registration in Australia and New Zealand are collated separately from the central government statistics agencies, however the coding schemes are closely related to mortality coding.

Unless otherwise noted, mortality data were coded to the International Classification of Diseases (ICD). Coding for lung cancer is discussed in Section 2.1.4.

Unless otherwise specified, data were available as cross-tabulations by single calendar year, by sex and by five year age groups through to ages 85 years and over.

2.1.1 WHO Mortality Database

For inter-country comparisons, mortality data were taken from the World Health Organisation (WHO) Mortality Database. The form of this database was a set of flat files organised by the different versions of the International Classification of Disease. Population data and coding schemes were also available.

The database was downloaded for files dated 19 July 2000. Although the Database is well documented and provides extensive data on deaths and populations, the complex set of flat files precludes easy use.

The WHO database provided data on mortality and populations for Australia and New Zealand for 1950–1995.

2.1.2 Australian data

As noted in Section 2.1.1, mortality and population data for 1950–1995 were available from the WHO Mortality Database. Australian mortality data by year of registration of death were obtained from the Australian Bureau of Statistics for 1996–1999 (Australian Bureau of Statistics, 2000). Lung cancer registrations for Australia were only available for the period since 1983 (Australian Institute of Health and Welfare and Au
1999) and have not been used here. Mortality data for cancer of the respiratory system for the period 1930–1949 were available from Holman and Armstrong (1982) by five year calendar periods.

Population data for 1930–1949 and 1996–1999 were obtained from the Australian Bureau of Statistics.

2.1.3 New Zealand data

As noted in Section 2.1.1, mortality and population data for 1950–1995 were available from the WHO Mortality Database. For New Zealand during 1996–1999, data on deaths and populations were obtained from the New Zealand Health Information Service (2000b). Lung cancer registrations for 1948–1996 were obtained from the New Zealand Health In
(2000a).

Population data for 1996–1999 were obtained from Statistics New Zealand. Statistics New Zealand also provided *de jure* populations from the New Zealand Census of Population and Dwellings for 1926, 1936, 1945 and for every five years between 1951 and 1996. These populations were available by single year age groups and by sex. All of the Census dates occurred during March for the given Census year.

2.1.4 Historical coding of lung cancer

Sites typically associated with cancer of the respiratory system include the trachea, bronchus, lung, pleura and mediastinum. As classifications have changed over time, it is difficult to obtain a consistent time series for "lung cancer" mortality or incidence.

Until very recently, registrations for malignant neoplasms in Australia and New Zealand were coded to the same classification system as mortality based on the topography or site of the neoplasm. The introduction of the International Classification of Diseases for Oncology in the different state cancer registries during the 1990s provided a cancer incidence classification system based on both topography and morphology. Usefully, it is possible to translate from the two-dimensional classification to the one-dimensional classification for most neoplasms. As a result, coding for mortality and incidence registration are usually comparable.

In Australia for the period 1930–1999 there were seven different mortality coding classifications. The fourth and fifth revisions for the Manual of the International List of Causes of Disease were used during 1930–1949, while the seventh through tenth revisions of the International Classification of Diseases were used during 1950–1999.

Two broad approaches have been used in the literature to resolve the problem of obtaining a longer-term time series. First, an inclusive definition has been used, such as for cancer of the respiratory system or of the intra-thoracic organs. Second, restricted definitions including particular sub-sites have also been used. For the period 1930–1999, two proposed time series are given in Table 2.1.

		Respiratory	Trachea, lung
Revision	Period[a]	system	and bronchus
Fourth	1930–1939	47	47b[b]
Fifth	1940–1949	47	47b[b]
Sixth	1950–1957	162–164	162–163[c]
Seventh	1958–1967	162–164	162–163[c]
Eighth	1968–1978	162–163	162
Ninth	1979–1998	162–165	162
Tenth	1999	C33–C39	C33-C34

Table 2.1: Proposed time series for lung cancer mortality for 1930–1999

Before the 1970s, almost all cancers of the respiratory system were for lung and bronchus. There have typically been very small numbers of cancers of the mediastinum and trachea. However cancer of the pleura has been rising in recent years due, in part, to asbestos exposure (Xu et al., 1985).

Consequently, the "lung cancer" time series is proposed as being the best for prediction, while the time series for "cancer of the respiratory system" will be adequate for a description of longer-term changes.

The longer-term time series may be affected by under-diagnosis. Table A.2 on

[a] Period used in Australia

[b] Includes pleura and excludes trachea

[c] Includes pleura

page 257 suggests that there has been considerable under-diagnosis of lung cancer deaths in the early to middle parts of the twentieth century. Under-diagnosis is related to age, where "the conventional wisdom that the accuracy of death certification is lower in the very elderly" applies (Coleman et al., 1993). This suggests caution in the interpretation of the earlier trends, particularly for older age groups.

However the mortality rate modelling performed in Chapter 10 has the oldest cohort reaching 70–74 years of age in 1995–1999, so that under-diagnosis is not expected to affect later modelling. The report in Appendix A attempts to resolve this issue by restriction to those aged 30–74 years and a period adjustment for under-diagnosis across those ages.

2.1.5 Standard populations

Age-standardised lung cancer rates have been standardised to Segi's World Population (as modified by Doll et al., 1966).

Prevalence data have been age-standardised to the 1991 Australian population. This was done because estimates from prevalence surveys are highly dependent upon the sampling frame, so comparisons between countries are unlikely to be valid. Moreover, Segi's World population is dissimilar to the current Australian population, being considerably younger.

2.1.6 Vitals by one year cohorts

Using available data and standard demographic techniques, the population and all cause mortality rates were estimated for single year of age and single calendar year by sex for Australia and New Zealand. Reliable estimates of the mid-point for open age groups were not available, so estimates for open age groups were not calculated. Moreover, given that mortality outcomes for infants were difficult to take into account, estimates for those aged under five years were also not calculated.

The New Zealand census populations were estimated using cumulative differencing across ages for the cohorts of interest (Shyrock et al., 1976). Thin plate splines were used to smooth the cumulative sums employing `proc tpspline` in `SAS` (with smoothing parameter $\lambda_0 = 0.1$). Then, for each cohort, the spline linear interpolation was calculated for each mid-year between each adjacent pairs of Censuses using similar smoothing parameters (Shyrock et al., 1976). Death data were available from 1898 to 1949 by five year age groups up to 80 years by single calendar years. Death rates for five year age groups were calculated using the numbers of deaths and aggregates of the population data. Logs of the death rates by single years of age were estimated by spline interpolation using the mid-points of the age groups.

Mortality and population data for Australia were available by five year age groups by single calendar years back to 1907. Logs of the death rates and population numbers by single year of age were estimated using spline interpolation using the mid-points of the age groups.

Estimates using linear interpolation for a cohort gave very similar results to the two-dimensional smoothing.

2.2 Tobacco consumption

Statistics for tobacco products available for consumption for Australia and New Zealand are collected from Customs and Excise (Department of Statistics and Department of H 1992; Ministry of Health, 2001; Winstanley et al., 1995). These data are available quarterly by source (customs, excise) and by type of tobacco product (manufactured cigarettes, cigars including cigarillos and cheroots, snuff, and loose tobacco).

There is no reason to believe that all products available for consumption will be consumed. These estimates do not include smuggled tobacco products ("chop-chop") that potentially account for a moderate proportion of total consumption.

Since 1939 in Australia, over 99% of tobacco consumption has been loose tobacco or manufactured cigarettes. Loose tobacco typically is used for both hand-rolled cigarettes and for pipe tobacco. The greater part of fine-cut tobacco in Australia since 1980 has been used for hand-rolled cigarettes, however consistent information on fine-cut tobacco is not available for earlier years.

There is a small gap in the Australian time series for 1915 when Customs and Excise information were not collected.

2.3 Smoking behaviour

Inclusion criteria for other smoking data sources were:

- For Australia or New Zealand
- Nationally representative
- One of:
  - Multiple cross-sectional surveys with the same sampling frame
  - Questions on time of smoking uptake and cessation.

The eligible surveys are described in Table 2.2.

Australian National Drug Strategy surveys prior to 1998, national health surveys for Australia and New Zealand national health surveys prior to 1996/97 were not included as they only provided smoking status at interview and did not have consistent sampling frames over time.

2.3.1 Data imputation

The methods used for data cleaning were comparable to those described by Harris (1983). Briefly, for individuals who reported being current smokers where their age of initiation was missing, the age of initiation was taken as the mean age of initiation for their ten year sex-birth cohort (1910–1919, 1920–1929, . . . , 1960–1969). For males, the mean ages by birth cohort were (18, 18, 18, 18, 17, 17). For females, the mean ages by birth cohort were (24, 22, 21, 19, 18, 17).

Again in keeping with (Harris, 1983), respondents who reported being former smokers who had a missing age of cessation were assumed to have smoked up until the time of interview.

2.3.2 Survey weighting

Although design information was available for most of the studies (excluding Anti-Cancer Council of Victoria surveys and AC Nielsen surveys prior to 1990), suitable survey analysis routines for most of the analyses were not available. Inclusion of the census results in a survey analytic framework was another technical difficulty.

As an approximate solution, effective cell sizes were used as weights in an analysis assuming independence. The effective cell sizes were calculated by scaling the sample weights so that they summed to the number of respondents divided by an approximate design effect.

In practice, the design effect varied between studies, age groups and sex, and by study question. For the multi-stage stratified surveys, the design effects tended to be less than two, while the design effects for the other study designs are expected to be less than two. As a conservative choice, a design effect of 2.0 has been assumed throughout.

See Section 3.6.5 for further discussion.

Country	Survey	Year(s)	Sampling frame	Sample size	Response rate
Australia	ACCV[a] smoking prevalence surveys	1974, 1976, 1980, 1983, 1986, 1989, 1992, 1995	Multistage cluster sample	∼8000 per year	∼50%
	NHFA[b] Risk Factor Prevalence Study	1980, 1983, 1989	Electoral rolls for capital cities, 25–64 years (except 1989: 20-69 years)	5617 (1980), 7640 (1983), 9309 (1989), total=22,566	75.9% (1980), 75.3% (1983), 74.7% (1989)
	National Drugs Strategy Household Survey	1998	Multistage cluster samples (three different samples)	10,030 (4012 for first sample)	55% (first sample)
New Zealand	Census of Population and Dwellings	1976, 1981, 1996	Census	2,201,178 (1976), 2,296,704 (1986), 2,786,220 (1996)	96.8% (1976), 98.1% (1981), 92.1% (1996), (item response)
	AC Nielsen smoking prevalence surveys	1983–2000	Multistage cluster sample	∼10,000 per year	∼50%
	National Health Survey	1996/97	Multistage cluster sample	7862	73.8%

Table 2.2: Summary of selected smoking surveys for Australia and New Zealand

[a] ACCV: Anti-Cancer Council of Victoria.

[b] NHFA: National Heart Foundation of Australia.

2.3.3 Anti-Cancer Council of Victoria prevalence surveys

Australian data on smoking prevalence have been collated since 1974 by the Anti-Cancer Council of Victoria (ACCV) from an omnibus market research survey performed by Roy Morgan Research. The definition of smoking used by the ACCV includes “regular” smoking of any tobacco products, rather than being restricted to cigarette smoking alone (see Table 2.3 for the question). This has little impact upon prevalence for females, as cigars and pipes account for a negligible proportion of all smokers. However over the period 1974–1995 a few percent of Australian male adults used cigars and pipes exclusively, which will inflate the male estimates relative to estimates for cigarette smoking alone.

The sampling frame for the ACCV data changed after 1995, after which the ACCV piloted a new smoking question that included responses for “occasional” smoking (Mullins et al., 2000). Although some validation was performed, the change in sampling frame and declining response rates effectively ended the time series in 1995.

Which one of those statements best describes you? Please say your answer and its number.

Smokers	
I smoke *only* cigarettes or roll-your-owns	1
I smoke cigarettes or roll-your-owns *and* also cigars or a pipe	2
I smoke cigars regularly, and *used to* smoke cigarettes	3
I smoke a pipe regularly, and *used to* smoke cigarettes	4
I smoke cigars regularly, and have *never* smoked cigarettes	5
I smoke a pipe regularly, and have *never* smoked cigarettes	6
Ex-smokers	
I used to smoke regularly, but *only* cigarettes	7
I used to smoke cigarettes regularly *and* also cigars or a pipe	8
I used to smoke *only* cigars or a pipe regularly, but not cigarettes	9
Never smokers	
I have *never* smoked at all	10

Table 2.3: ACCV smoking question, 1974–1995

2.3.4 Risk Factor Prevalence Study

The National Heart Foundation of Australia Risk Factor Prevalence Study was a series of surveys carried out in 1980, 1983 and 1989 (Risk Factor Prevalence Study Management C 1990; Bennett and Magnus, 1994). Respondents were a systematic random sample from the Federal electoral rolls for seven study centres in major urban areas. Electoral enrolment is compulsory for Australian citizens aged 18 years and over. Each survey was held during May to November of each year.

The questionnaires included questions about cigarette brands and switching to lower tar cigarettes (see Table 2.4). More importantly for this analysis, questions were also asked about age of smoking initiation and date of smoking cessation.

Data were provided by the Social Sciences Data Archive.

Q26. Have you ever smoked cigarettes, cigars or pipes regularly?

Yes .... □

No .... □ → Go to Question 33

Q27. At what age did you start smoking regularly?

I started smoking at □□ years of age

Q28. Have you given up smoking?

Yes, I gave up smoking in .... □□/19□□

No, I still smoke .... □

If you have given up smoking please answer the following question:

Q29. How much did you smoke?

I used to smoke □□ manufactured cigarettes on a working day
□□ manufactured cigarettes on a leisure day
□□□ grams "hand-rolled" cigarette tobacco per week
□□ cigars per week
□□□ grams pipe tobacco per week.

**If you CURRENTLY SMOKE please answer Questions 30 to 32;
otherwise go to Question 33.**

Q30. How much do you smoke?

I currently smoke □□ manufactured cigarettes on a working day
□□ manufactured cigarettes on a leisure day
□□□ grams "hand-rolled" cigarette tobacco per week
□□ cigars per week
□□□ grams pipe tobacco per week.

*NOTE: a $1\frac{3}{4}$ ounce pouch of cigarette tobacco equals 50 grams

Q31. Which brand of manufactured cigarettes do you usually smoke?
(Copy the name from a packet if possible)

I don't smoke manufactured cigarettes .... □

The brand I usually smoke is .... □

Q32. Have you switched to lower tar manufactured cigarettes?

Yes, in .... □□/19□□

No .... □

I don't know .... □

Table 2.4: Smoking questions from the 1989 Risk Factor Prevalence Survey

2.3.5 1998 National Drugs Strategy Household Survey

The 1998 National Drugs Strategy Household Survey was commissioned by the Australian Institute of Health and Welfare and performed by Roy Morgan Research (Australian Institute of Health and Welfare, 1999). The multi-stage stratified design used a probabilistic design, randomly selecting individuals from a randomly selected household. Information was collected using a combination of personal interview and self-completed questionnaire.

For the parts of the questionnaire related to use of tobacco, see Tables 2.5 and 2.6.

Three different sampling frames were employed. Formal analysis could have been restricted to the first sample ($n = 4012$), as the sampling frame was more rigorous. All three of the samples were included for the purposes of increased sample size, however caution is suggested for any interpretation.

Data were provided by the Social Sciences Data Archive, with additional design information provided by the Australian Institute of Health and Welfare.

2.3.6 NZ Census of Population and Dwellings

Questions relating to smoking behaviour were asked in the New Zealand Census of Population and Dwellings during 23 March 1976, 24 March 1981 and 5 March 1996. Questions were asked of those aged 15 years and over about their current and past regular cigarette smoking, where regular smoking was defined as smoking one or more tobacco cigarettes per day (see Table 2.7). The questions were similar in form to the questions asked in the 1996/97 New Zealand Health Survey.

Data for 1976 and 1981 were available by five year age groups by sex (Department of Statistic 1979, 1983). Data for 1996 were available by single year of age by sex and were provided by Statistics New Zealand.

Prevalence of a particular smoking category was estimated by the number in that category divided by the number of respondents. Non-respondents were assumed to be non-differentially distributed between the smoking categories.

The proportion of respondents who did not give a specified response to the Census smoking question has changed over time. In 1976, 3.2% of adults did not give a specified response compared with 1.9% in 1981 and 7.9% in 1996. An adjustment for differential non-response by post-stratification for age, sex and ethnic group (European, NZ Māori, Pacific Islands, Asian, Other and Unspecified) did not alter age-specific or total smoking prevalence.

During the 1976 and 1981 Censuses, respondents were also asked to specify the

G1. About what proportion of your friends and acquaintances smoke tobacco?
CROSS ONE BOX ONLY

All .... □

Most .... □

About half .... □

A few .... □

None .... □

G2. In the last 12 months, have you or any other members of this household regularly smoked tobacco in the home?
REGULARLY SMOKED MEANS AT LEAST ONE CIGARETTE, CIGAR, OR PIPE A DAY

Yes inside the home .... □

No, only smoke outside the home .... □

No-one at home regularly smokes .... □

G3. Have you personally ever tried smoking cigarettes or other forms of tobacco?

Yes .... □

No .... □ → Next Section.

G4. Have you ever smoked a full cigarette?

Yes .... □

No .... □ → Next Section.

G5. About what age were you when you smoked your first full cigarette?
ENTER WHOLE YEARS ONLY (E.G. 21, 35, 47)

Age in years .... □□

G6. Who supplied you with your first cigarette?
CROSS ONE BOX ONLY

Friend or acquaintance .... □

Sibling (brother or sister) .... □

Parent .... □

Spouse/partner .... □

Other relative .... □

Stole it .... □

Purchased it myself from shop/tobacco retailer .... □

Other .... □

Can't recall .... □

G7. Would you have smoked at least 100 cigarettes (manufactured or roll your own), or the equivalent amount of tobacco in your life?

Yes .... □

No .... □

G8. Have you ever smoked on a daily basis? Yes □

No .... □ → G12.

G9. About what age were you when you started smoking daily?
ENTER WHOLE YEARS ONLY (E.G. 21, 35, 47)

Age in years .... □□

G10. Are you still a daily smoker?

Yes .... □ → G12.

No .... □

G11. About what age were you when you last smoked daily?
ENTER WHOLE YEARS ONLY (E.G. 21, 35, 47)

Age in years .... □□

Table 2.5: 1998 National Drugs Strategy Household Survey smoking questions (part a)

G12. In the last 12 months, have you . . .
CROSS AS MANY BOXES AS APPLY

Successfully given up smoking (for more than a month)? □

Tried to give up unsuccessfully? □

Changed to cigarette brand with lower tar or nicotine content? □

Reduced the amount of tobacco you smoke in a day? □

None of these □

G13. Where do you usually obtain your cigarettes now?
CROSS ONE BOX ONLY

Friend or acquaintance □

Sibling (brother or sister) □

Parent □

Spouse/partner □

Other relative □

Steal them □

Purchase from shop/tobacco retailer □

Other □

Not relevant – don't smoke now □ → G16.

G14. Are you planning on giving up smoking?

Yes, within 30 days □

Yes, after 30 days, but within the next □

3 months □

Yes, but not within the next 3 months □

No □

G15. Please read through all the statements below, and then cross the one statement which best describes your current use of tobacco/cigarettes.

Now smoke occasionally, **but less than once a week** □

Now smoke occasionally, **but at least once a week, about** . . .

5 or less cigarettes a week □

6 – 10 cigarettes a week □

11 – 15 cigarettes a week □

16 – 20 cigarettes a week □

21 – 25 cigarettes a week □

26 – 30 cigarettes a week □

31 or more a week □

Now smoke regularly, **everyday or most days**, about . . .

5 or less cigarettes a day □

6 – 10 cigarettes a day □

11 – 15 cigarettes a day □

16 – 20 cigarettes a day □

21 – 25 cigarettes a day □

26 – 30 cigarettes a day □

31 or more a day □

IF STILL SMOKE: □ → Next Section.

G16. About what age were you when you last smoked tobacco?
ENTER WHOLE YEARS ONLY (E.G. 21, 35, 47)

Age in years □□

Table 2.6: 1998 National Drugs Strategy Household Survey smoking questions (part b)

Q27. Do you smoke cigarettes regularly (that is, one or more per day)?
COUNT ONLY tobacco cigarettes
DON'T COUNT pipes, cigars or cigarillos
yes .... □→ Go to Q29
no .... □

Q28. Have you ever been a regular smoker of one or more cigarettes per day?
yes .... □
no .... □

(Help note for Question 27: Count tailor-made and roll-your-owns but only if they are tobacco cigarettes; do NOT count herb or marijuana cigarettes)

Table 2.7: 1996 New Zealand Census smoking questions

number of cigarettes smoked on the day prior to the Census.

2.3.7 AC Nielsen surveys

AC Nielsen surveys have been used as the primary time series for smoking prevalence in New Zealand since 1983 (Ministry of Health, 2000). The questions were included in a general omnibus market research survey. Data were available by five year age groups up to 64 years, 65–74 years, and 75 years and over. Information on former smokers was not available.

Either or both of the categories for cigarette smoking from the questionnaire (see Table 2.8) were used to indicate cigarette use.

2.3.8 1996/97 NZ Health Survey

Recall of changes in smoking behaviour was sought in the 1996/97 New Zealand Health Survey (NZHS) (Ministry of Health, 1999b). The sample frame was non-institutionalised New Zealand adults aged 15 years and over. The respondents were interviewed face to face during the period October 1996 to October 1997. The ASCII data file included question responses, sample weights, together with survey design information on the primary sampling units and sampling strata. The design effect for the prevalence of smoking in the total population was 1.83. Younger people and males tended to have higher design effects, with a range of 0.91 to 2.88.

After recoding, 12 respondents were excluded due to missing values. As age of interview, initiation and cessation were recorded by whole years, an adjustment for duration of smoking was made if two or more of these events occurred in the same

Which of these products, if any, do you EVER smoke?

Ready-made cigarettes .... □

Roll-your-own cigarettes .... □

Cigars/Cigarillos .... □

Pipe .... □

[None] .... □→ End

[IF EVER SMOKE READY-MADE CIGARETTES]

On average how many ready-made cigarettes do you smoke on an AVERAGE DAY? .... □□

[IF EVER SMOKE ROLL-YOUR-OWN CIGARETTES]

On average how many roll-your-own cigarettes do you smoke on an AVERAGE DAY? .... □□

Table 2.8: AC Nielsen omnibus survey

year.

2.3.9 Comparison of smoking data sources

The differences in sampling frames and variations in the questions preclude direct comparisons of prevalence between the different surveys.

Definitions for Australia have tended to emphasise smoking of tobacco products. For the historical time series for 1974–1995 collated by ACCV, respondents were asked whether they smoke cigarettes or roll-your-owns without asking about regularity, while they were asked whether they regularly smoke other tobacco products (Table 2.3). The Risk Factor Prevalence Study asked about regular use of tobacco products, while the 1998 National Drugs Strategy Household Survey asked about daily smoking.

The 2001 Australian National Health Data Dictionary specifies that tobacco smoking status should be collected for daily, weekly, irregular smokers, ex-smokers and those that never smoked (Australian Institute of Health and Welfare, 2001). Usefully, ex-smokers are defined in terms of whether an individual has smoked 100 or more cigarettes or equivalent in their lifetime. This definition has been validated against the ACCV time series, where smoking from ACCV gives similar results to combining daily and weekly smokers from the National Health Data Dictionary (Mullins et al., 2000). Whether this equates to "regular" smoking from other data sources is moot.

Q51. Do you smoke one or more tobacco cigarettes a day?

yes □

no □ → Q54

don't know □ → Q54

Q52. About how many cigarettes do you smoke on an average day?

1 to 10 a day? □

11 to 20 a day? □

21 to 30 a day? □

31 or more a day? □

don't know □

Q53. Which of the following best describes you <u>now</u>? (tick only one)

I have no thoughts of quitting smoking □ → Q57

I think I need to consider quitting someday □ → Q57

I think I should quit smoking but I'm not quite ready □ → Q57

I think about doing things that will help me quit smoking □ → Q57

I'm doing things that will help me quit smoking □ → Q57

don't know □ → Q57

Q54. Generally, if someone has been smoking cigarettes near you, how would you say you find the smoke?

enjoyable on the whole □

does not bother me □

bothers me slightly □

bothers me a lot □

don't know □

Q55. Have you ever been a regular smoker of one or more cigarettes per day?

yes □

no □ → Q58

don't know □ → Q58

Q56. At what age did you last regularly smoke one or more cigarettes per day?

age □□

don't know □

Q57. At what age did you start regularly smoking one or more cigarettes per day?

age □□

don't know □

Q58. Does anyone (including yourself) smoke cigarettes inside your home every day or most days?

yes □

no □

don't know □

Table 2.9: 1996/97 New Zealand Health Survey smoking questions

The New Zealand surveys have tended to emphasise regular cigarette smoking. The AC Nielsen surveys asked about “ever use” for different tobacco products so that contemporaneous prevalence estimates for current cigarette smoking were 1–2% higher than estimates from other New Zealand surveys.

Chapter 3

Multi-state smoking model: formulation and theory

Abstract

A theoretical development is presented for a multi-state model of smoking in a population. The model includes states for never, current and former smokers, with transitions for uptake, cessation, recall error by former smokers as never smokers and differential mortality. Estimation is described for two different forms of data: for *current status* data based on respondents stating their current state of smoking; and for *retrospective* data which are based on recall of smoking behaviour up to and including the time of interview. Some technical aspects related to model fitting are discussed.

3.1 Introduction

As seen from Chapter 1, the prevalence of current smoking is a useful measure for population-based smoking behaviour. However, smoking prevalence describes smoking at a point in time, while there is also interest in how smoking changes over time. This thesis proposes that a dynamic description of population smoking using multi-state models will have a number of important benefits.

First, multi-state models provide an integrated framework for estimation of a range of smoking parameters. Such parameters include the distribution of smoking duration by current and former smokers, time since smoking cessation by former smokers, cumulative measures of dose and survival estimates based on smoking status at a prior time. One application discussed in Chapter 1 is predicting the effect of a tobacco related intervention for diseases which have a long time course and are related to cumulative smoking parameters.

Second, estimation of the smoking dynamics leads to a better understanding of the system. As an example, bounds can be found on how a system has been known to change in the past, which suggests bounds on how the system may change in the future. Model simplification may be suggested, for example uptake may be negligible after a certain age, simplifying any analytical development. Alternatively, model complexity may be found, such as the need to account for recall error by former smokers as never smokers (see Chapter 6).

The models may help to provide more valid estimates for historical smoking prevalence. Moreover, it may be possible to make model-based predictions of future or past smoking prevalence. This second approach has been used by Mendez and colleagues to argue that US smoking prevalence will continue to decline and that US Federal smoking targets are unlikely to be achieved (Mendez et al., 1998; Mendez and Warner, 2000). However their projections were based on arguable assumptions, reinforcing the view that projections should be undertaken with care (see Appendix B).

There are certain limitations to a dynamic modelling approach, particularly due to data availability. The majority of smoking data are available from complex surveys. Longitudinal data are rarely available and would be expensive to collect. Changes are slow however and recall may provide much of the necessary information at a population level.

Estimation of the dynamics is more onerous than estimation of prevalence. There is a responsibility to ensure that the model has a basis in reality and that model inputs, and hence model predictions, are validated (Gunning-Schepers, 1999).

3.1.1 Outline of the chapter

In this chapter, a literature review will provide some context to multi-state models in epidemiology. Then the class of dynamic smoking models will be described. The models will then be formally expressed and possible extensions outlined. The models are identified as being multi-state models. Multi-state models are discussed in general and then some specific results for the smoking model are considered. Estimation is then considered for two different forms of data: *current status* data based on respondents stating their current state of smoking; and *retrospective* data which are based on recall of smoking behaviour up to and including the time of interview. Finally, some other technical aspects related to model fitting are discussed.

3.2 Literature review

To provide some context for the use of multi-state models, various model classes will be reviewed, the relationship between mathematical models and estimation outlined, and some applications of multi-state models will be described.

3.2.1 Model classes

Multi-state models broadly include many of the models that are fundamental to epidemiology. In their simplest form, *cohort studies* follow an individual from study inclusion to the occurrence of an event (Breslow and Day, 1987). States include the entry state (usually "healthy") and the state described by the event, such as death or the incidence of cancer. This epidemiological design uses pervasively the concepts of *rates* which is the expected number of events per *person-year* of exposure for a given *risk set.* Person-years of exposure and risk sets can be understood graphically by the movement of an individual across the *Lexis diagram.* For an historical review and statistical development, see Keiding (1990).

A closely related model, important in both epidemiology and demography, is the single decrement life table, where there are states for being alive and for death. For an epidemiological review, see Estève et al. (1994) and see the classic monograph by Shyrock et al. (1976) for demographic methods.

Another common multi-state model is the three state illness-death model, where individuals can move from a healthy state to death or to a sickness state. From the sickness state, an individual can move to the death state, or in some models may be able to recover to the healthy state again. An obvious complication is that *duration* in the sick state may affect the likelihood of either death or failure to recover. An important association is how the *prevalence* of those in the sickness state relates to the *incidence* of new cases to the sickness state. For an analytical development, see Chiang (1968). Keiding (1991) provides a broad and inclusive review of more recent developments that have influenced the methodological approach taken here.

The illness-death model is an example of a multi-state model which has *competing risks*, where there can be different types of events from a given state, and individuals could have *multiple events.*

More general multi-state models have been receiving increasing attention. Useful reviews include Hougaard (1999) and Commenges (1999). The development of multi-state models used here borrows from both of these references, together with the encyclopaedic monograph by Andersen et al. (1993).

Although multi-state models subsume most of the population-based models in epidemiology, there has been some interest in modelling for interactions between individuals or modelling for complex health states for each individual. This general class of models, called *micro-simulation*, has only recently become available given increased computing power. It is gaining some popularity for modelling population health (Wolfson, 1994) and for modelling social systems, as used at the Australian National Centre for Social and Economic Modelling (Walker, 2000). Mention should also be given to models for continuous risk functions that possibly deserve more attention (Tolley and Manton, 1991).

3.2.2 Mathematical models and estimation

Multi-state models require an understanding of both the underlying mathematical models, methods for estimation of the parameters, and data with which to estimate the parameters. Possibly stating the obvious, poor model specification or poor estimation will limit the validity of any conclusions. In population health, relatively little may be known about a system, so that the range of mathematical models that can be developed will be constrained by the range of valid estimates.

The functional form of the transition intensities may not be well established, limiting the application of parametric models. This suggests the need for non-parametric estimation, which would involve the use of modern survival and smoothing methods. Where there is evidence for proportionality of the intensities between groups, semi-parametric estimation may be used.

Estimation may also require the use of different data sources with different sources of variability. As an example, a model may include data for mortality rate ratios for different exposures from aetiological studies, total mortality from vital statistics, and prevalence at different points in time from survey data. Valid point and interval estimation then requires stronger model assumptions or more sophisticated estimation methods.

3.2.3 Applications

3.2.3.1 Smoking

There have been several previous smoking applications of methods related to multi-state modelling. For a cohort-based analysis of smoking, Harris (1983) introduced the use of synthetic cohort reconstruction based on retrospective data. Harris noted that prevalence estimates were biased due to differential survival and proposed a suitable adjustment. A natural progression of this research is to investigate the associated component changes. Earlier research reported cumulative measures of change (e.g. Christie et al., 1986), with some limited recent efforts to estimate the cessation rates themselves (Burns et al., 1997b). Birkett (1997) presented a proportional hazards analysis for comparing smoking rates between groups, however no effort was made to establish proportionality of the rates between the groups. The analysis of retrospective data in Chapter 5 is closely related to these developments, with extensions using modern hazard estimation methods and adjustments for differential mortality using the full multi-state model.

For an analysis using data based on current smoking status, Mendez et al. (1998) fitted a simplified dynamic smoking model using non-linear least squares to the prevalence of current smokers from the U.S. National Health Interview Surveys. An extension of this approach is used in Chapter 7, modelling for both current and former smokers and using a likelihood function.

For a hypothetical simulation of smoking, Hakulinen and Pukkula (1981) used a multi-state model starting from baseline values for smoking prevalence and duration and then considered various scenarios for uptake and cessation. Cessation alternatives were 0%, 10% and 20% for a five year period. The authors also took account of differential mortality for all causes and for lung cancer, including parameters for dose and duration of smoking. The all cause mortality model used in Chapter 4 makes different assumptions, where condition-specific rate ratios are assumed constant between populations. The form of the risk function is also different from those fitted in Chapter 10.

Tolley et al. (1991) used a multi-state Markov smoking model with states for age and duration of current smoking and duration of cessation. The Markov model included transitions between all states and accounted for competing causes of death. Estimates for initiation and cessation rates were empirically based, however there was limited reporting of how the estimates were derived. The theoretical model development was presented in Tolley and Manton (1991). The multi-state smoking model developed in this chapter has fewer states, however the estimation methods used here are more sophisticated.

Interestingly, both Hakulinen and Pukkula (1981) and Tolley et al. (1991) used the smoking exposure estimates to forecast lung cancer mortality rates. These developments are closely related to the smoking models developed in this chapter.

3.2.3.2 Carcinogenesis

Multi-state models subsume probably the most popular mathematical models for carcinogenesis. The multi-stage model of carcinogenesis involves a finite set of heritable changes in a single cell line moving in a fixed sequence towards a pre-clinical cancer (Armitage and Doll, 1954). This is equivalent to a multi-state model with each point in the sequence being a state with transition rates moving along the sequence of states. The transition rates are potentially dependent upon environmental conditions, such as smoking history. This model is equivalent to birth processes as discussed by Chiang (1968).

Two other popular models for carcinogenesis are the clonal growth model (Gaffney and Altsh 1988) and two-mutation recessive oncogenesis model (Moolgavkar et al., 1989). These models involve two stages and include different assumptions about cell growth and differentiation at the different states. Cell growth can be represented either by a multi-state model with a large number of states or by a model with a small number of states and a feedback mechanism. The latter form lies outside of the analysis presented later in this chapter.

These models are the basis for the lung cancer risk functions developed in Chapter 10, where they will be discussed in more depth.

3.2.3.3 Cancer incidence, survival and mortality

The illness-death model is suited to modelling cancer prevalence taking account of incidence, survival and general mortality. This is technically complicated as the mortality rate for those with cancer is dependent upon both their place on the Lexis diagram and on duration of having cancer.

Multi-state model estimation has been carried out by Verdecchia et al. (2001) and Gail et al. (1999). Inputs to this estimation process include relative survival, where survival is adjusted for the population mortality rate.

3.2.3.4 Infectious diseases

Multi-state models provide a flexible framework to represent the mechanism for a disease. This may explain the popularity of such models for representing infectious diseases, where their study has reached a certain level of sophistication (Anderson and May, 1992). One common multi-state model has states for those in the population who are susceptible, infected and immune. The modelling is intended to describe parameters that offer insight on the probability of endemic years and on controlling the spread of disease.

Recent important applications include back projections for the HIV/AIDS epidemic (De Angelis et al., 1998) and modelling the infection of cows with bovine spongiform encephalopathy and incident cases of variant Creutzfeldt-Jakob Disease in humans (Donnelly and Ferguson, 1999). The applications tend to include data from disparate sources, with models that have a demographic component and a disease mechanism. Fitting of the models has used a variety of techniques, including full Bayesian models, maximum likelihood estimation for full dynamic models and modelling of incidence using serial prevalence surveys (Becker and Marschner, 2001).

The associated likelihood estimation methods were influential in the analysis of the census data presented in Chapter 6. Moreover, the analysis of serial prevalence survey data was influential in the theoretical development of the analysis of the current status survey data in Chapter 7.

3.2.3.5 Other health applications

There has been recent interest in using multi-state models to model the relationship between incidence and prevalence. The Global Burden of Disease Study requires these parameters to estimate burden due to years of life lost to disability (Murray and Lopez, 1997b) and has developed software (`DISMOD`) for this purpose. A variety of other health applications for multi-state models are discussed in Andersen et al. (1993) and Commenges (1999).

3.2.3.6 Applications outside of health

Applications of multi-state models in health borrow strongly from previous applications in demography and ecology, where both disciplines are concerned with changes in *populations*. As an example of this cross-fertilisation, Anderson and May (1992) had previously been involved in ecological research. These two disciplines provide a rich source of methodology and forms of applications for health researchers. For some demographic applications, see Hougaard (1999).

3.3 Smoking models

In this section, I describe a multi-state smoking model and introduce some notation. Possible model extensions are also discussed. A formal development is postponed until Section 3.4.

A detailed model for smoking behaviour is shown in Figure 3.1. The model has a series of states (newborn, migrant, never, former and current smoking, dead) and a set of transitions between the states (birth, smoking uptake, smoking cessation, smoking restarting, recall error of a former smoker as a never smoker, migration and death). In the following, former smokers and ex-smokers are taken as being equivalent.

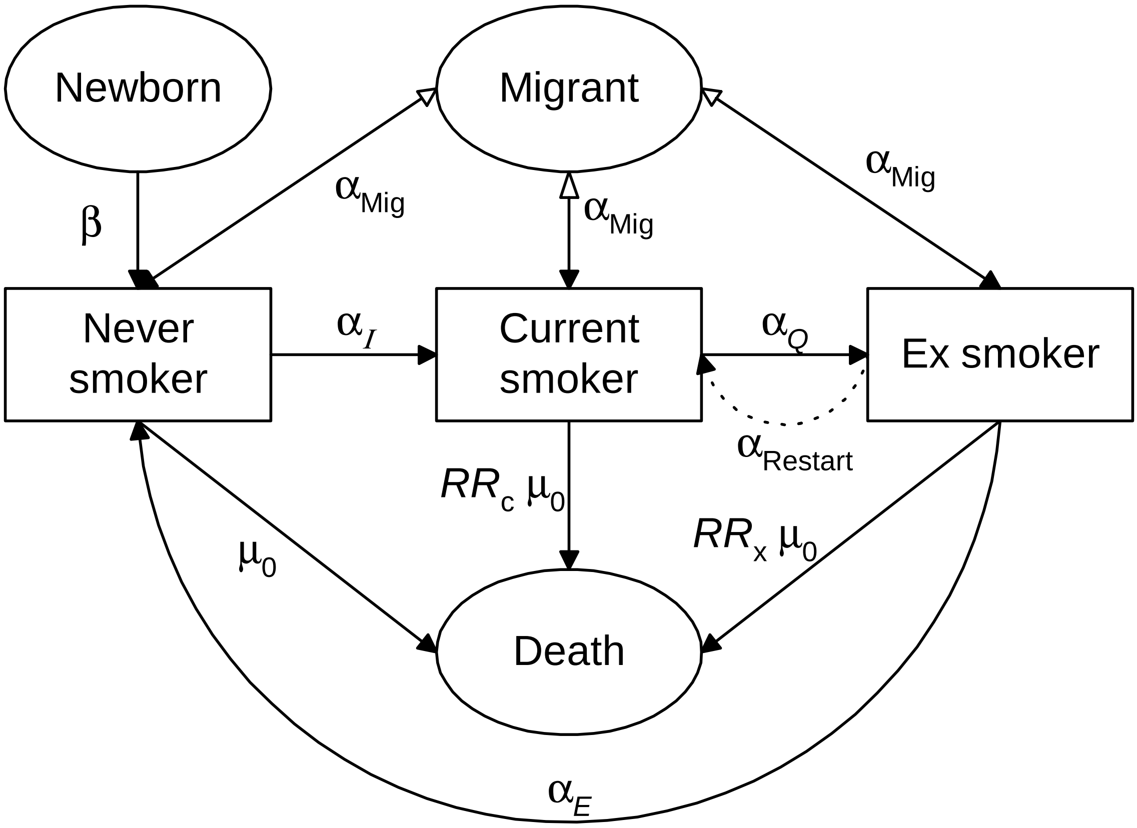


Figure 3.1: Graphical representation of a detailed smoking model

At a point in time, the population can be described by the size of the population states (never, current and former smokers), or the prevalence of each population state. Changes over time are generally described by the initial values for the states and either the probabilities or rates of changing from one state to another.

The symbols on Figure 3.1 represent rates, where β is the birth rate and α_{Mig} is the net rate of immigration expressed in terms of the resident population. The mortality rates for the total population and for never smokers are represented by μ and μ_0, respectively. All cause mortality rate ratios for current and former smokers compared with never smokers are represented by RR_c and RR_x, respectively, so that the mortality rates for current and former smokers will be $RR_c\mu_0$ and $RR_x\mu_0$, respectively. The net smoking initiation rates, smoking quit rates and former smoker recall error rate are represented by α_I, α_Q and α_E, respectively. Finally, the rate of former smokers restarting smoking is represented by α_{Restart}. The numbers of never, current and ex- smokers are represented by the upper case letters of N, C and X, respectively. The total population size is represented by T $(= N + C + X)$. The prevalence of never, current and former smokers is represented by π_n $(= N/T)$, π_c and π_x, respectively.

3.3.1 Some tractable models

At this point, the rates have not been well defined. Moreover, some simplification of the model would make estimation of the different parameters more practical. A more practical model (labelled Model 1) is presented in Figure 3.2. Here, migration has been ignored, with the implicit assumption that migration will not affect estimates of smoking prevalence. A proof of this is given in Section 3.5.6. Moreover, net cessation is considered, where the result of quitting and restarting is summarised by one rate. Finally, births are implicitly accounted for by population counts.

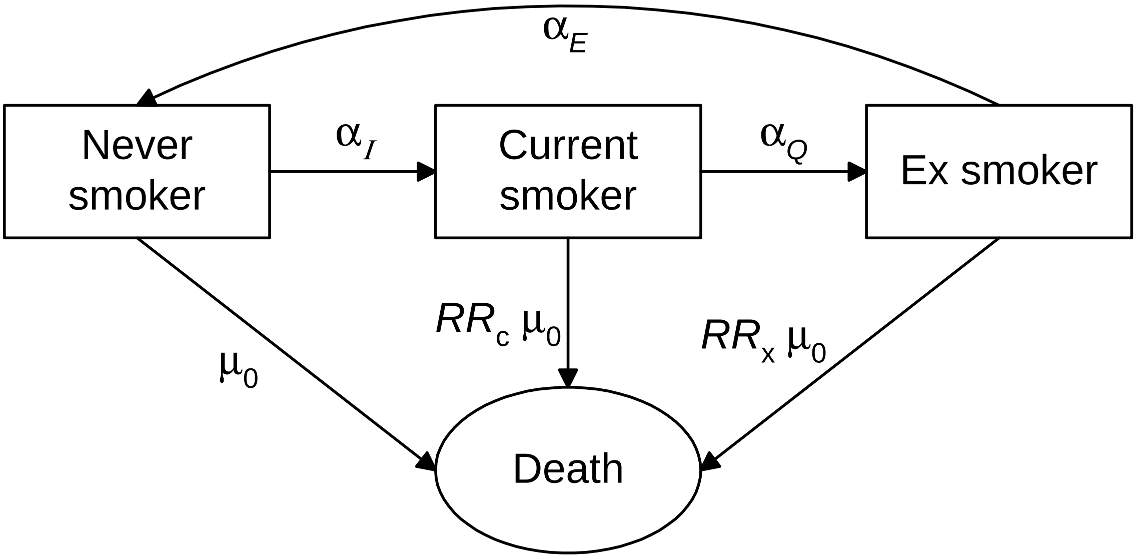


Figure 3.2: Graphical representation of a smoking model with uptake, cessation, recall error and differential mortality (Model 1)

An important sub-model is the closed population with only initiation and cessation, ignoring death (Model 2: see Figure 3.3). One example is for a population sample where respondents are asked whether and when they have started or stopped smoking (see Chapter 5). The sample is conditional upon survival and reports how respondents recall their past behaviour based on today's perception.

There are several other common sub-models of Model 1. For youth smoking, mortality can be ignored, leaving a cycle of transitions. When recall error is assumed negligible, we have a system that does not have a cycle and is more amenable to analytic expressions for transition probabilities.

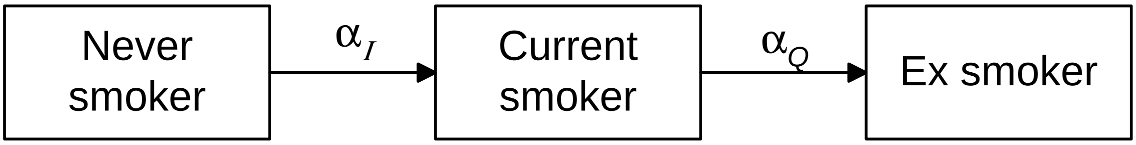


Figure 3.3: Graphical representation of a smoking model with initiation and cessation (Model 2)

3.3.2 Smoking classifications and transitions

A variety of data sources will provide information on smoking (see Chapter 2). In particular, survey data and aetiological data will be combined. For this reason, the smoking states and smoking transitions must be defined in quite a general manner. *Current smokers* will thus be defined as those in the population who at a point in time self-identify as being either a current, regular, or daily smoker. *Never smokers* and *former smokers* are those who self-identify as having never smoked or having been a current smoker, respectively.

Definitions of smoking initiation and cessation are complicated by possible short-term changes, such as those due to smoking experimentation or short-term quit attempts. Following Mendez et al. (1998), the definition adopted here is for *net* smoking initiation and cessation. This technically measures an aggregate change, which for an individual may be loosely interpreted as a definite change in behaviour: either becoming an established current smoker, or having “successfully” quit. For the retrospective analysis, having successfully quit is taken as having quit for three or more years.

A further complication for smoking cessation is that the time distribution may have spikes immediately following a large tobacco control campaign or other publicity. Estimation of these spikes would require precise and frequent surveys. Moreover, the cessation rate would be expected to decline below the average level after a spike because of “cessation harvesting”, where those smokers who were candidates for cessation would have been provided a catalyst for quitting and were no longer available to quit. In practice, it is useful to assume that the cessation rate is a smooth function over time.

Recall error or misclassification by former smokers who later identify as never smokers may be appreciable, as suggested by van de Mheen and Gunning-Schepers (1994). The implication is that at older ages there may be a number of self-identified never smokers who had in fact been current smokers at some time. From an aetiological perspective, the bias from this is unclear, as those “never smokers” may be more likely to have smoked less and for a shorter period of time than the self-identified former smokers (Lee and Forey, 1996). Using Census data, recall error rates are estimated in Chapter 6.

One extension to Model 1, suggested by Figure 3.1, would be to include a rate of re-starting smoking due to "unsuccessful" quitting. This extension would be useful for its ability to model the effect of an intervention that increases the number attempting to quit or the length of the average quit attempt. Given that the rate of re-starting would be dependent upon time since cessation, the model would not be Markov (see Section 3.4.2) and would require a different form of model compared with that proposed in this chapter.

Another extension of some merit would be to include an additional state for former smokers who have changed their recall to be identified as never smokers. This would allow modelling the mortality rate as being intermediate between former and never smokers and quantification of the number with the recall error. The model is also intuitive in the sense that there would not be a cycle in the model, where former smokers who have recall error would usually not be candidates for smoking initiation, as they are in Model 1. The main disadvantage of this model is that initiation rates are usually measured in terms of self-identified never smokers, which includes misclassified former smokers. Therefore estimation of initiation rates in terms of true never smokers cannot easily be measured.

3.3.3 Mortality

Estimates of total mortality are available from vital statistics. These statistics are generally assumed to be reasonably precise and valid. Given estimates of the all cause mortality rate ratios and prevalence for current and former smokers compared with never smokers, the total mortality rate for a group can be divided by smoking status using the equation

$$\mu = \pi_n \mu_0 + \pi_c RR_c \mu_0 + \pi_x RR_x \mu_0$$

which can be manipulated to allow the estimation of the baseline (never smoker) mortality rate μ_0 and hence mortality rates by smoking status using

$$\mu_0 = \frac{\mu}{\pi_c(RR_c - 1) + \pi_x(RR_x - 1) + 1} \tag{3.1}$$

There are a variety of sources of all-cause mortality rate ratios for current and former smokers by age and sex. Issues include stability of rate ratios between populations and over time. Importantly, survival estimates become increasingly sensitive to the rate ratio estimates at older ages (see Section 3.5.5). Estimates of the mortality rate ratios will be considered in more depth in Chapter 4.

One natural extension to the multi-state model would be to include absorbing states for different causes of death. Two possible advantages of such an approach would be for estimating attributable mortality and adjusting for competing risks. For examples using multiple causes of death, see Tolley et al. (1991) and Tolley and Manton (1991). The latter example uses the mortality rates in the likelihood function with a multi-state model incorporating smoking cessation and multiple causes of death.

For an analytical development, functional forms for mortality can be introduced, as used by Anderson and May (1992). However there have been appreciable changes in mortality over the last fifty years, so that the following analysis will use observed mortality.

3.3.4 Births

Births could be included in the model in several ways. First, observed births can be incorporated into the model, either implicitly by using population estimates or explicitly using the observed or predicted birth rates. Second, for an analytical development using stochastic processes, births can be assumed to follow a Poisson process (Brillinger, 1986; Keiding, 1991).

3.3.5 Migration

Migration has been assumed to be non-differential based on smoking status. Non-differential migration and small differences in smoking patterns will not affect proportional estimates (see Section 3.5.6). Moreover, rates for lung cancer in New South Wales for 1973–1998 were not different among those born outside Australia compared with those born in Australia (Goumas et al., 2001), suggesting that migrant smoking patterns were similar to the residential population. One possible extension would be to use census data for a full population demographic analysis, including numbers for births, deaths and migration (Shyrock et al., 1976), performing a sensitivity analysis for differential migration rates.

Note that the formulation used for migration in the detailed smoking model is different from that suggested by Chiang (1968), who suggested that migration should be represented as absolute numbers over time rather than as a rate with respect to some population. The given formulation was chosen for its convenience.

3.3.6 Duration

Duration of current smoking and of time since cessation can be included in one of two ways. First, the model can be assumed to be Markov (see Section 3.4.2), so that the duration distributions can be estimated from the fully specified model. This is the approach followed later in the thesis. Second, the model can be extended to include states for duration of current smoking and duration as a former smoker, with systematic transitions in the duration states as a person ages. This approach has been used successfully by Tolley et al. (1991), although the transition rates were assumed to be homogeneous over time, which may not be valid. A useful extension would be to investigate which set of assumptions is more valid. Such an investigation would be limited by the need for a large data set.

3.3.7 The changing cigarette

An important aspect of modelling historical smoking exposure is accounting for the changing cigarette and the changing pattern of smoking use (see Chapter 1 and Wilkenfeld et al., 2000). Tar and nicotine delivery following burning and "inhalation" in a standard manner have been declining over time, however measurement of the bio-available content is difficult. Moreover, given lower nicotine cigarettes in more recent years, smokers may have altered their smoking behaviour to obtain higher levels of nicotine.

There is limited historical information on changes in the composition of cigarettes. This issue is discussed further in Chapter 9.

3.3.8 Heterogeneity of exposure

Throughout the development of the model, it has been assumed that the population considered is homogeneous. However, there is significant variation in exposure to smoking over time within most populations. As a simple example, considering smoking exposure for a group that contains both males and females would fail to show the marked differences in smoking history between the two gender groups. Given that smoking is associated with most socio-economic variables (Ministry of Health, 1999b), there is expected to be significant heterogeneity in any estimates derived.

Variation in dose and duration can be incorporated in the analysis. However care is required where there may be other important sources of variability that have not been explained. One possible extension would be to describe sub-populations more precisely by the use of covariates or finer stratification, however any such investigation would be limited by the available data.

3.4 Multi-state models

To outline the following development, the machinery for multi-state models is defined. The main models of interest are multi-state models with piecewise homogeneous Markov transitions. Two important relationships due to Chapman and Kolmogorov are presented, which provide a mechanism to model the multi-state model. Throughout the development it has been assumed that regularity conditions are satisfied, such that the transition rates are assumed bounded. For a complete theoretical development, see Andersen et al. (1993).

3.4.1 Definitions

A multi-state model with J different states can be represented by the random variable $U(t)$, which takes the value $j,\ j = 1, \ldots, J$ when in state j at time t. The *history* $\mathcal{F}(t)$ is generated by $\{U(u), u \leq t\}$, being all of the information about the process up to a time t. The history includes the set $\{N(t), T_1, \ldots, T_{N(t)}, U(0), \ldots, U(T_{N(t)})\}$ where $N(t)$ is the number of transitions up to time t and $T_m,\ m = 1, \ldots, N(t)$ is the time of the m^{th} transition.

The changes in state can be described by either the *transition probabilities* $P_{jk}(s,t)$, the probability of moving from state j to state k between times s and t, where

$$P_{jk}(s,t) = P(U(t) = k \mid U(s) = j;\ \mathcal{F}(s))$$

or by the *transition intensities* $\alpha_{jk}(t, \mathcal{F}(t-))$ which can be defined heuristically as

$$\alpha_{jk}(t, \mathcal{F}(t-)) = \lim_{\mathrm{d}t \to 0} P_{jk}(t, t + \mathrm{d}t)/\mathrm{d}t \quad (j \neq k),$$

where $\mathcal{F}(t-)$ is the history immediately before time t. From this heuristic definition, $\alpha_{jk}\,\mathrm{d}t$ can be interpreted as the probability of moving from state j to state k over the small period $\mathrm{d}t$. It is useful to define $\alpha_{jj} = -\sum_{k \neq j} \alpha_{jk}$. Moreover, synonyms for transition intensities include *rates* and *hazards*.

Following the development by Commenges (1999), the development for one process can be applied carefully to a population. In particular, there is a need to consider and describe heterogeneity in a population. Taking a sample of I individuals, who have a set of parameters (or characteristics) $\boldsymbol{Z}_i$, where $i = 1, \ldots, I$, the intensity function can then be described by

$$\alpha^i_{jk}(t, \mathcal{F}(t-)) = \alpha_{jk}(t, \boldsymbol{Z}_i, \mathcal{F}(t-)).$$

This assumes that the α_{jk} are homogeneous conditional upon $\boldsymbol{Z}_i$, which may be, at best, an approximation. Given that this is often tacitly assumed, the onus is upon the investigator to find groups that are at least largely homogeneous.

Let the $P_{jk}(s,t)$ form the *transition probability matrix* $\boldsymbol{P}(s,t) = (P_{jk}(s,t))$ and the $\alpha_{jk}(t)$ form the *transition intensity matrix* $\boldsymbol{\alpha}(t) = (\alpha_{jk}(t))$. Moreover, let the cumulative intensity be defined as

$$A_{jk}(t) = \int_0^t \alpha_{jk}(s)\,\mathrm{d}s \quad (j \neq k)$$

and let $A_{jj} = -\sum_{k\neq j} A_{jk}$. Then define the *cumulative intensity matrix* as $\boldsymbol{A} = (A_{jk})$.

Finally, $S_i(s,t)$ represent the probability of *survival* from state i at time s to time t, where survival is well-defined for a closed system where the absorbing states represent mortality, such that

$$S_i(s,t) = \sum_{j\in\mathcal{L}} P_{ij}(s,t)$$

where $\mathcal{L}$ is the set of live states.

3.4.2 Markov and semi-Markov models

An important class of multi-state models assumes that the history of the process can be ignored — the *Markov* models. For these models, the transition rate can be represented by $\alpha_{jk}(t, \mathcal{F}(t-)) = \alpha_{jk}(t)$. Similarly for the transition probabilities

$$P_{jk}(s,t) = P(U(t) = k \mid U(s) = j).$$

(which is not dependent upon $\mathcal{F}$ except for the state $U(s)$ at time s). The transition intensities can be assumed dependent upon time t and also upon covariates.

A common alternative to the Markov property is that the intensity is dependent only upon the time in the current state (rather than being dependent on the current state and the main time scale t). These models form the *semi-Markov* models, with transition intensity

$$\alpha_{jk}(t, \mathcal{F}(t-)) = \alpha_{jk}(t - T_{N(t-)}).$$

3.4.3 Two important relationships

For Markov models, an important relationship due to Chapman and Kolmogorov is that

$$P_{jk}(s,t) = \sum_{m \in \mathcal{S}} P_{jm}(s,s_1)P_{mk}(s_1,t)$$

for $s < s_1 < t$, where $\mathcal{S}$ is the set of states. This can be expressed in matrix-form as

$$\boldsymbol{P}(s,t) = \boldsymbol{P}(s,s_1)\boldsymbol{P}(s_1,t). \tag{3.2}$$

which is independent of the intermediate time s_1. This provides a simple mechanism for calculating transition probabilities over a period of time given transition probability matrices for a partition of the period: simply multiply the transition probability matrices together.

Another important relationship, due to Kolmogorov, is the forward differential equation

$$\left.\begin{array}{rcl} \boldsymbol{P}(s,s) & = & \mathbf{I} \\ \dfrac{\partial \boldsymbol{P}(s,t)}{\partial t} & = & \boldsymbol{P}(s,t)\boldsymbol{\alpha}(t). \end{array}\right\} \tag{3.3}$$

where $\mathbf{I}$ is the identity matrix (Andersen et al., 1993). The solution to this differential equation provides a mechanism for calculating analytically or numerically the transition probability matrix given the transition intensity matrix. An equivalent form for a continuous process that is more useful for the purposes of estimation, involves the product-integral (bold $\prod$) [1] in the manner of Andersen et al. (1993). If $\boldsymbol{A}$ is integrable, which is trivially satisfied if (α_{jk}) are continuous and bounded, then

$$\boldsymbol{P} = \prod_{(t,u]} [\mathbf{I} - \mathrm{d}\boldsymbol{A}(u)] \tag{3.4}$$

This formulation is more general than the Kolmogorov forward differential equations. However, all of the intensity functions of interest will satisfy Equation (3.3). This formulation is essentially a generalisation of the univariate analytical development that yields the Kaplan-Meier analysis (see Section 3.6.2.2). The formulation has been introduced here because it offers a bridge between the piecewise homoge-

[1] The use of a script capital $\boldsymbol{\pi}$ for denoting the product integral used by Andersen et al. (1993), which is attributed to Gill and Johansen (1990), is non-standard and not readily available (in LaTeX).

neous approach suggested in the next section with a continuous time approach. The former approach is useful for incorporating data sources that are discrete in time, such as for relative survival or for life tables. The latter approach can be used for estimation when event history data with exact times are available.

3.4.4 Piecewise homogeneous Markov models

For a time homogeneous Markov model, Equation (3.3) can be solved for $\boldsymbol{P}(s,t)$ by *matrix exponentiation*, where

$$\boldsymbol{P}(s,t) = \exp((t-s)\boldsymbol{\alpha})$$

The form of Kolmogorov's equation and its analytical solution is similar to a common univariate rate equation (Andersen et al., 1993). Taking Equations (3.2) and (3.3) together, we have

$$\begin{aligned}\boldsymbol{P}(t,u) &= \prod_{(t_i,t_{i+1}]\text{: interval of }(t,u]} \boldsymbol{P}(t_i,t_{i+1}) \\ &= \prod_{(t_i,t_{i+1}]\text{: interval of }(t,u]} \exp\left[(t_{i+1}-t_i)\cdot\boldsymbol{\alpha}(t_i,t_{i+1})\right].\end{aligned}$$

3.4.5 Time scales

The development to this point has only used one time scale. However different time scales can be used. Examples of common times scales include age, calendar time and time since study entry.

For a population development, it is often useful to describe changes in rates across the *Lexis diagram*, which is a Cartesian co-ordinate system with axes for calendar time t and age a. A leading diagonal of the Lexis diagram follows a birth cohort $c = t - a$ through time.

For a full specification of the transition intensities, we would have a functional form $\alpha(a,t)$. For notational convenience, as there is a linear dependency between a and t for a given cohort c, it is often useful to only represent one of the time scales.

For a technical discussion of time scales, see Andersen et al. (1993, Chapter 10). Keiding (1990) provides a historical and statistical review for Lexis diagrams.

3.5 Model developments

The following presents some results that are used in the analysis of the model.

Analytical approximations to Model 1 are provided in Appendix 3.A on page 90. In practice, the use of Chapman-Kolmogorov and Kolmogorov equations provided a more flexible mechanism for numerical analysis.

3.5.1 Transition intensity matrices

For Model 1, let the state vector be ordered by never, current and former smokers, and finally with deaths. Then $\boldsymbol{\alpha}$ is represented by

$$\boldsymbol{\alpha} = \begin{bmatrix} -\mu_0 - \alpha_I & \alpha_I & 0 & \mu_0 \\ 0 & -RR_c\mu_0 - \alpha_Q & \alpha_Q & RR_c\mu_0 \\ \alpha_E & 0 & -RR_x\mu_0 - \alpha_E & RR_x\mu_0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

It is often convenient, both analytically and numerically, to restrict the analysis to the live states, so that

$$\boldsymbol{\alpha} = \begin{bmatrix} -\mu_0 - \alpha_I & \alpha_I & 0 \\ 0 & -RR_c\mu_0 - \alpha_Q & \alpha_Q \\ \alpha_E & 0 & -RR_x\mu_0 - \alpha_E \end{bmatrix}. \tag{3.5}$$

3.5.2 Differential equations

For the developments of methods based on cross-sectional prevalence data, it is useful to have an alternative formulation of the multi-state models as a set of partial differential equations. For Model 1, the differential equations are

$$\frac{\partial N}{\partial t} + \frac{\partial N}{\partial a} = -\left[\mu_0 + \alpha_I\right] N + \alpha_E X \tag{3.6}$$

$$\frac{\partial C}{\partial t} + \frac{\partial C}{\partial a} = -\left[RR_c\mu_0 + \alpha_Q\right] C + \alpha_I N \tag{3.7}$$

$$\frac{\partial X}{\partial t} + \frac{\partial X}{\partial a} = -\left[RR_x\mu_0 + \alpha_E\right] X + \alpha_Q C \tag{3.8}$$

where the demographic and epidemiological parameters are considered to also be functions of age a and time t. Through most of the following, partial derivatives have been used to be clear that changes by age and by time are being considered, which is equivalent to following a cohort c through time. As birth cohort is constant for an individual (a person can only be born on one day), then, under the conditions of the multivariable Chain Rule, the following operator relationship holds:

$$\frac{\partial}{\partial a} + \frac{\partial}{\partial t} = \frac{\mathrm{d}}{\mathrm{d}t}. \tag{3.9}$$

3.5.3 Duration distributions

Expressions can be found from this mathematical model for the distribution of duration of current smoking for current and for former smokers, and the distribution of times since cessation for former smokers. These parameters will be used for regression modelling of lung cancer mortality rates from vital statistics, hence age is a natural time scale for analysis.

For this development, we consider a given birth cohort c at age a. The probability of being a current smoker and having a duration of smoking $\mathrm{Dur}_{\mathrm{sm}}$ between duration d_1 and duration d_2 is

$$P(d_1 \leq \mathrm{Dur}_{\mathrm{sm}} \leq d_2 \,|\, \mathrm{age}\, a, \mathrm{cohort}\, c) = \int_{a-d_2}^{a-d_1} P_{nn}(0,u)\alpha_I(u)P_{cc}(u,a)\ \mathrm{d}u$$

The mean for the duration of smoking by current smokers $\overline{\mathrm{Dur}}_{\mathrm{sm|curr}}$ at age a is constructed by integrating over all ages, conditioning on being a current smoker, so that

$$\begin{aligned}\overline{\mathrm{Dur}}_{\mathrm{sm|curr}} &= \int_0^a \mathrm{P}(\text{Started at age } u \,|\, \text{current at age } a)(a-u)\ \mathrm{d}u \\ &= \left[\int_0^a P_{nn}(0,u)\alpha_I(u)P_{cc}(u,a)(a-u)\ \mathrm{d}u\right] / P_{nc}(0,a).\end{aligned}$$

The variance of the duration of current smoking can be estimated using a similar formulation:

$$\begin{aligned}\mathrm{var}(\mathrm{Dur}_{\mathrm{sm|curr}}) &= \left[\int_0^a P_{nn}(0,u)\alpha_I(u)P_{cc}(u,a)(a-u)^2\ \mathrm{d}u\right] / P_{nc}(0,a) \\ &\quad - \overline{\mathrm{Dur}}^2_{\mathrm{sm|curr}}.\end{aligned}$$

For the case of homogeneous rates, a mid-point approximation can be used. For those aged a completed years (that is, for those in the age interval $[a, a+1)$ years)

$$
\begin{aligned}
\overline{\mathrm{Dur}}_{\mathrm{sm|curr}} \approx & \left\{ \sum_{i=0}^{a-1} \left[(a-i)\frac{P_{nn}(0,i)+P_{nn}(0,i+1)}{2}\alpha_I(i+1/2) \right.\right. \\
& \left. \times \frac{P_{cc}(i,a)+P_{cc}(i+1,a)+P_{cc}(i,a+1)+P_{cc}(i+1,a+1)}{4} \right] \\
& \left. + \frac{1}{3}\frac{2P_{nn}(0,a)+P_{nn}(0,a+1)}{3}\alpha_I(a+1/2)P_{cc}(a,a+1)^{1/3}\frac{1}{3} \right\} \\
& / \frac{P_{nc}(0,a)+P_{nc}(0,a+1)}{2}. \qquad (3.10)
\end{aligned}
$$

The third line of the equation accounts for smoking initiation during the final age interval. By assuming for the interval that the probability distributions for smoking initiation and for the population distribution are uniform and independent, then the average smoking initiation is one-third through the interval. Moreover, the average observation for those that began smoking would be two thirds through the interval, while the average smoking duration and the interval for integration would be one third of a year (hence the $\frac{1}{3}$ at the beginning and the end of the third line). A similar argument suggests the weighting of P_{nn} and P_{cc} for that interval.

A similar development was used for the duration of time since cessation, such that

$$\overline{\mathrm{Dur}}_{\mathrm{cess|former}} = \left[\int_0^a P_{nc}(0,u)\alpha_Q(u)P_{xx}(u,a)(a-u)\ \mathrm{d}u \right] \ / \ P_{nx}(0,a).$$

For duration of smoking for current smoking, the integration must be over time of initiation and time of cessation.

$$
\begin{aligned}
\overline{\mathrm{Dur}}_{\mathrm{sm|former}} &= \int_0^a \int_u^a \mathrm{P}(\text{Started at age } u \text{ and quit at age } v \,|\, \text{former at age } a) \\
&\quad \times (v-u)\ \mathrm{d}v\ \mathrm{d}u \\
&= \left[\int_0^a \int_u^a P_{nn}(0,u)\alpha_I(u)P_{cc}(u,v)\alpha_Q(v) \right. \\
&\quad \left. \times\ P_{xx}(v,a)(v-u)\ \mathrm{d}v\ \mathrm{d}u \right] \ / \ P_{nx}(0,a)
\end{aligned}
$$

The development of the mid-point approximation took account of the following cases:

- an interval including initiation, cessation *and* observation (ignored, as contribution would be negligible)
- an interval including initiation and cessation (duration=1/3 with adjustment of probabilities similar to Equation (3.10))
- an interval including cessation and observation (duration=mid-point separation-1/6 with adjustment of probabilities similar to Equation (3.10)).

For those aged a completed years, the mid-point approximation will be

$$\begin{aligned}
\overline{\mathrm{Dur}}_{\mathrm{sm|former}} \approx\ & \Bigg\{ \sum_{i=0}^{a-1}\sum_{j=i}^{a-1} \Bigg[(j-i)\frac{P_{nn}(0,i)+P_{nn}(0,i+1)}{2}\alpha_I(i+1/2)\alpha_Q(j+1/2) \\
& \times \frac{P_{cc}(i,j)+P_{cc}(i+1,j)+P_{cc}(i,j+1)+P_{cc}(i+1,j+1)}{4} \\
& \times \frac{P_{xx}(j,a)+P_{xx}(j+1,a)+P_{xx}(j,a+1)+P_{xx}(j+1,a+1)}{4} \Bigg] \\
& + \frac{1}{9}\sum_{i=0}^{a-1} \Bigg[\frac{2P_{nn}(0,i)+P_{nn}(0,i+1)}{3}\alpha_I(i+1/2)\alpha_Q(i+1/2) \\
& \times P_{cc}(i,i+1)^{1/3} \\
& \times \frac{P_{xx}(i,a)+2P_{xx}(i+1,a)+P_{xx}(i,a+1)+2P_{xx}(i+1,a+1)}{6} \Bigg] \\
& + \frac{1}{3}\sum_{i=0}^{a-1} \Bigg[(a-i-1/6)\frac{P_{nn}(0,i)+P_{nn}(0,i+1)}{2}\alpha_I(i+1/2)\alpha_Q(i+1/2) \\
& \times \frac{2P_{cc}(i,a)+2P_{cc}(i+1,a)+P_{cc}(i,a+1)+P_{cc}(i+1,a+1)}{6} \\
& \times P_{xx}(a,a+1)^{1/3} \Bigg] \Bigg\} \\
& \Big/ \ \frac{P_{nx}(0,a)+P_{nx}(0,a+1)}{2}. \qquad (3.11)
\end{aligned}$$

3.5.4 Dose and consumption

Different measures of cigarette consumption are available from different sources. For the following, *dose* will be defined as the rate of cigarette consumption *per smoker*. An alternative measure of consumption is the rate of cigarette consumption *per person*.

Cumulative measures are required for dose for current and former smokers and for the consumption rate per person. For a given measure of dose, let dose and cumulative dose be represented by "Dose" and "CumDose", respectively. Moreover, let consumption rate per person and cumulative consumption per person be represented by "Cons" and "CumCons", respectively.

The main assumption in the following development is that the *average* dose for those smokers at time u who survive to time t will be similar to those smokers who do not survive.

The mean dose per current smoker at age a is

$$\begin{aligned}\overline{\text{Dose}}_{\text{curr}} &= \int_0^a P(\text{Started at time } u \,|\, \text{current at age } a) \\ &\quad \times (\text{Average dose from } u \text{ to } a)\, \mathrm{d}u \\ &= \left[\int_0^a P_{nn}(0,u)\,\alpha_I(u)\,P_{cc}(u,a)\int_u^a \frac{\overline{\text{Dose}}(v)}{a-u}\,\mathrm{d}v\,\mathrm{d}u\right]/P_{nc}(0,a)\end{aligned}$$

with variance

$$\text{var}(\text{Dose}_{\text{curr}}) = \left[\int_0^a P_{nn}(0,u)\,\alpha_I\,P_{cc}(u,a)\int_u^a \frac{\text{var}(\text{Dose})(v)}{(a-u)^2}\,\mathrm{d}v\,\mathrm{d}u\right]/P_{nc}(0,a).$$

In practice, the integral can be calculated using a mid-point approximation. The average dose per former smoker is

$$\begin{aligned}\overline{\text{Dose}}_{\text{former}} &= \left[\int_0^a\int_u^a P_{nn}(0,u)\,\alpha_I(u)\,P_{cc}(u,v)\,\alpha_Q(v)\,P_{xx}(v,a)\right. \\ &\quad \left.\times \int_u^v \frac{\overline{\text{Dose}}(w)}{v-u}\,\mathrm{d}w\,\mathrm{d}v\,\mathrm{d}u\right]/P_{nx}(0,a)\end{aligned}$$

with variance

$$\begin{aligned}\text{var}(\text{CumDose}_{\text{former}}) &= \left[\int_0^a\int_u^a P_{nn}(0,u)\,\alpha_I(u)\,P_{cc}(u,v)\,\alpha_Q(v)\,P_{xx}(v,a)\right. \\ &\quad \left.\times \int_u^v \frac{\text{var}(\text{Dose})(w)}{(v-u)^2}\,\mathrm{d}w\,\mathrm{d}v\,\mathrm{d}u\right]/P_{nx}(0,a).\end{aligned}$$

The mean cumulative dose per current smoker at age a is

$$\begin{aligned}\overline{\text{CumDose}}_{\text{curr}} &= \int_0^a P(\text{Started at time } u \,|\, \text{current at age } a) \\ &\quad \times (\text{Cum dose from } u \text{ to } a)\, \mathrm{d}u \\ &= \left[\int_0^a P_{nn}(0,u)\, \alpha_I(u)\, P_{cc}(u,a) \int_u^a \overline{\text{Dose}}(v)\, \mathrm{d}v\, \mathrm{d}u\right] / P_{nc}(0,a)\end{aligned}$$

with variance

$$\text{var}(\text{CumDose}_{\text{curr}}) = \left[\int_0^a P_{nn}(0,u)\, \alpha_I\, P_{cc}(u,a) \int_u^a \text{var}(\text{Dose})(v)\, \mathrm{d}v\, \mathrm{d}u\right] / P_{nc}(0,a).$$

In practice, the integral can be calculated using a mid-point approximation. The mean cumulative dose per former smoker is

$$\begin{aligned}\overline{\text{CumDose}}_{\text{former}} &= \left[\int_0^a \int_u^a P_{nn}(0,u)\, \alpha_I(u)\, P_{cc}(u,v)\, \alpha_Q(v)\, P_{xx}(v,a)\right. \\ &\quad \left. \times \int_u^v \overline{\text{Dose}}(w)\, \mathrm{d}w\, \mathrm{d}v\, \mathrm{d}u\right] / P_{nx}(0,a)\end{aligned}$$

with variance

$$\begin{aligned}\text{var}(\text{CumDose}_{\text{former}}) &= \left[\int_0^a \int_u^a P_{nn}(0,u)\, \alpha_I(u)\, P_{cc}(u,v)\, \alpha_Q(v)\, P_{xx}(v,a)\right. \\ &\quad \left. \times \int_u^v \text{var}(\text{Dose})(w)\, \mathrm{d}w\, \mathrm{d}v\, \mathrm{d}u\right] / P_{nx}(0,a).\end{aligned}$$

Finally for this section, cumulative consumption per person may be required for the surviving cohort. This is represented by

$$\overline{\text{CumCons}_T} = \int_0^a \overline{\text{Cons}}(u)\, S_c(u,a)/S_T(u,a)\, \mathrm{d}u$$

where $S_c(u,a)$ and $S_T(u,a)$ are survival estimates from age u to time a for current smokers and the total cohort, respectively.

If dose and consumption are expressed in similar units, then $\overline{\text{Cons}} = \overline{\text{Dose}}\, \pi_c$. Moreover, if dose and prevalence are assumed to be independent then for the product of two random variables (Mood et al., 1974), we have

$$\begin{aligned}\mathrm{var}(\mathrm{CumCons}_T) &= \int_0^a \left[\overline{\mathrm{Dose}}^2 \mathrm{var}(\pi_c) + \pi_c^2 \mathrm{var}(\mathrm{Dose}) + \mathrm{var}(\pi_c) + \mathrm{var}(\mathrm{Dose})\right] \\ &\quad \times S_c(u,a)/S_T(u,a)\,\mathrm{d}u.\end{aligned}$$

3.5.5 Sensitivity for rate ratios

It is useful to have some sense of how the transition probabilities are sensitive to the rate ratio estimates. For any such analysis, it is important to note that μ_0 is also a function of RR_c and RR_x. Taking partial derivatives of $\boldsymbol{\alpha}$ with respect to RR_c and RR_x, the second order approximation for $\boldsymbol{\alpha}$ given a change δ in the two rate ratios is

$$\begin{aligned}\boldsymbol{\alpha}(t \mid RR_c+\delta, RR_x+\delta) &\approx \boldsymbol{\alpha}(t \mid RR_c, RR_x) \\ &\quad + \delta\,\frac{\mu_0^2}{\mu}\,\mathrm{diag}\left(\begin{bmatrix} \pi_c+\pi_x \\ \pi_c+\pi_x-1-\pi_x(RR_x-RR_c) \\ \pi_c+\pi_x-1-\pi_c(RR_c-RR_x)\end{bmatrix}\right) \\ &\quad + \delta^2\,\frac{\mu_0^3}{\mu^2}\,\mathrm{diag}\left(\begin{bmatrix} -(\pi_c^2+2\pi_c\pi_x+\pi_x^2) \\ 2RR_x\pi_c^2+RR_c(\pi_x^2-\pi_c^2) \\ 2RR_c\pi_x^2+RR_x(\pi_c^2-\pi_x^2)\end{bmatrix}\right).\end{aligned}$$

where diag($\boldsymbol{x}$) is a diagonal matrix taking the values of the vector $\boldsymbol{x}$. If the diagonal elements are not appreciably different, then an approximate expression for the revised transition probability matrix from time t to time $t+1$ is

$$\begin{aligned}&\boldsymbol{P}(t,t+1 \mid RR_c+\delta, RR_x+\delta) \\ \approx\ &\boldsymbol{P}(t,t+1 \mid RR_c, RR_x) \\ &\times \mathrm{diag}\left\{\exp\left(\delta\,\frac{\mu_0^2}{\mu}\begin{bmatrix} \pi_c+\pi_x \\ \pi_c+\pi_x-1-\pi_x(RR_x-RR_c) \\ \pi_c+\pi_x-1-\pi_c(RR_c-RR_x)\end{bmatrix}\right)\right\}.\end{aligned}$$

This approximation can be applied for the period s to time t, such that

$$\begin{aligned} & \boldsymbol{P}(s,t \mid RR_c + \delta, RR_x + \delta) \\ \approx\ & \boldsymbol{P}(s,t \mid RR_c, RR_x) \\ & \times \operatorname{diag}\left\{ \exp\left(\delta \sum_{j=s}^{t-1} \frac{\mu_0^2}{\mu} \left[\begin{array}{c} \pi_c + \pi_x \\ \pi_c + \pi_x - 1 - \pi_x(RR_x - RR_c) \\ \pi_c + \pi_x - 1 - \pi_c(RR_c - RR_x) \end{array} \right] \right) \right\}. \end{aligned}$$

The estimated revised transition probabilities are scaled by a factor $\exp(\mu_0^2/\mu)$, which matches with the expectation that survival estimates will be most affected by errors in the rate ratios at older ages. Also the errors are cumulative, so that short-term estimates are likely to be little affected, while care is required for longer periods.

3.5.6 Effect of migration

Let the population net immigration rate, represented as a rate with respect to total population size, be represented by α_{Mig}. If migration is non-differential, so that $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \alpha_{Mig}\mathbf{I}$, then

$$\begin{aligned} \boldsymbol{P}(t, t+1) &= \exp(\boldsymbol{\alpha}) \\ &= \exp(\alpha_{Mig}\mathbf{I} + \boldsymbol{\alpha}_0) \\ &= \exp(\alpha_{Mig}) \exp(\boldsymbol{\alpha}_0) \end{aligned}$$

so that $\boldsymbol{P}$ can be estimated from transition rates which ignore migration and then scaled by $\exp(\alpha_{Mig})$. This will not affect proportional estimates, such as prevalence. For a longer period, the scale factor will be the exponential of the cumulative sum for annual migration.

If migration is differential, then let the migration transition rates be represented by $\operatorname{diag}(\boldsymbol{\alpha}_{Mig}) = \alpha_{Mig}\mathbf{I} + \operatorname{diag}(\boldsymbol{\delta}_{Mig})$, where $\operatorname{diag}(\boldsymbol{\delta}_{Mig})$ is a diagonal matrix with values which are the deviations from the population migration rate. Using a Taylor's series approximation (as $\exp(\mathbf{A} + \mathbf{B}) = \exp(\mathbf{A})\exp(\mathbf{B})$ if $\mathbf{AB} = \mathbf{BA}$)

$$\begin{aligned}
\boldsymbol{P}(t, t+1) &= \exp(\boldsymbol{\alpha}) \\
&= \exp(\boldsymbol{\alpha}_0 + \alpha_{Mig}\mathbf{I} + \mathrm{diag}(\boldsymbol{\delta}_{Mig})) \\
&\approx \exp(\alpha_{Mig})\exp(\boldsymbol{\alpha}_0) + (1 + 2\,\alpha_{Mig})\mathrm{diag}(\boldsymbol{\delta}_{Mig})
\end{aligned}$$

This provides a means to estimate the effect of differential migration on the (step) transition probability matrix. In general, for a moderate migration rate (α_{mig} being small on an absolute scale), then the transition probability matrix will be affected by small changes on the diagonal consistent with the departure from the average migration rate. The Taylor's series approximation effectively assumes no change during period for those migrating.

3.6 Design and analysis

The two general forms for data are retrospective data, where survey respondents recall whether and when they began and stopped smoking, and current status data, where respondents state whether they currently identify as being a current or former smoker. For retrospective data, the transition intensity can be estimated using survival estimation. For the current status data, the change in prevalence has a simple relationship with the transition intensities. These two approaches are discussed in the following sections.

3.6.1 Censoring and truncation

Event history data, such as that being used to measure smoking initiation and cessation, have some important characteristics. A respondent who is recorded as being a never smoker may potentially start smoking in the future, where the respondent is said to be *right censored* for smoking initiation. A respondent cannot begin to give up smoking until they have started smoking; their exposure to cessation requires smoking initiation. The exposure to cessation is said to be *left truncated.*

For current status data, those who have begun smoking are *left censored* with regards to smoking initiation, where we know that they started smoking prior to interview. Moreover, those respondents who have stopped smoking at time of interview are doubly left censored, having both started and stopped smoking prior to interview.

Conditions on the censoring and truncation are required to be satisfied for the following analyses to remain valid. The main assumption is that of *non-informative censoring*, where the reason that the respondent was censored is not related to the events of interest.

An introduction to censoring patterns is given in Klein and Moeschberger (1997). Issues relating to non-informative censoring, such as differential survival, together with possible approaches when the conditions are not satisfied, are reviewed by Keiding (1992).

3.6.2 Retrospective data

One consequence of the previous discussion is that estimates of the transition intensity matrix or the cumulative intensity matrix allow calculation of the transition probability matrix. In practice, the transition probability matrices and transition intensities can be estimated from the steps in the cumulative intensity function.

3.6.2.1 Available software

The dynamic modelling required software incorporating modern survival and smoothing methods. The obvious choice of software was thus `R/S-Plus`, however not all of the software required were available in a suitable form. Survival analysis for survey data can be performed using `SUDAAN` (`SURVIVAL` procedure) and `Fortran` code (`Coxreg`) from the National Cancer Institute (Korn et al., 1997). The latter code is able to handle left truncated data.

For hazard estimation of data that are only right censored, a variety of software are available. For `R/S-Plus`, packages are available using kernel based methods with boundaries corrections (`muhaz`), log-spline methods (`logspline`), local likelihood (`locfit`), and a less formal loess smoothing approach (`hazcov`). However these packages are restricted to data that are only right censored. Moreover, only some of these packages account for weighted data, and none are available for survey data. Recent `Fortran` code, `PHMPL`, is available for maximum penalised likelihood estimation of data that are left truncated and right censored or interval censored (Joly et al., 1999). However there is no facility to account for weighted data. An ideal solution would be to develop similar software that accounts for the survey design.

The absence of suitable “canned” software required development of suitable `R/S-Plus` code.

3.6.2.2 Aalen-Johansen estimator

To review the notation used for a population, we have a set of counting processes $\{X_i(t) : i = 1, \dots, I\}$ for the population represented by $N_{jk}(t)$. Moreover, let the number in state j at time t (the *risk set*) be denoted by $Y_j(t)$. Then

$$\hat{A}_{jk}(t) = \int_0^t \frac{1}{Y_j(s)} \mathrm{d}N_{jk}(s) \qquad (j \neq k)$$

and

$$\hat{A}_{jj}(t) = -\sum_{k \neq j} \hat{A}_{jk}.$$

Then Equation (3.4) gives the Aalen-Johansen estimator of the transition probability matrix:

$$\hat{\boldsymbol{P}} = \prod_{(t,u]} \left[\mathbf{I} - \mathrm{d}\hat{\boldsymbol{A}}(u) \right]. \tag{3.12}$$

In practice this is a simple finite product of matrices, where at the transitions

$$\mathrm{d}\hat{A}_{jk}(u) = \frac{\Delta N_{jk}(u)}{Y_j(u)},$$

which is the step in the cumulative intensity $\hat{A}_{jk}$, and zero otherwise.

Taking a simple example, consider a death process (1=alive, 2=dead), where $\alpha_{12} =$ mortality rate. Then $\Delta N_{12}(T_i)$ is the number of deaths at observed time T_i where $s < T_1 < T_2 < \cdots < T_m$. Then Equation (3.12) becomes

$$\begin{aligned} \hat{\boldsymbol{P}} &= \prod_i \begin{bmatrix} 1 - \dfrac{\Delta N_{12}(T_i)}{Y_1(T_i)} & \dfrac{\Delta N_{12}(T_i)}{Y_1(T_i)} \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} \prod_i \left(1 - \dfrac{\Delta N_{12}(T_i)}{Y_1(T_i)}\right) & 1 - \prod_i \left(1 - \dfrac{\Delta N_{12}(T_i)}{Y_1(T_i)}\right) \\ 0 & 1 \end{bmatrix} \end{aligned}$$

where the top-left cell of the matrix $(= \hat{P}_{11}(s,t))$, can be recognised as the Kaplan-Meier product limit survival estimator.

3.6.2.3 Non-parametric estimation of the transition intensity

Smoothing of the steps in the cumulative intensity function is required to estimate the transition intensity. A popular approach is to use kernel functions, as suggested by Ramlau-Hansen (1983):

$$\begin{aligned}\hat{\alpha}(t) &= b^{-1}\int K\left(\frac{t-s}{b}\right)\mathrm{d}\hat{A}(s)\\ &= b^{-1}\sum_i K\left(\frac{t-T_i}{b}\right)\frac{\Delta N(T_i)}{Y(T_i)}\end{aligned} \tag{3.13}$$

where K is a kernel function such that it is non-negative with $\int K = 1$. The variance can be estimated by

$$\begin{aligned}\sigma^2[\hat{\alpha}(t)] &= b^{-2}\int\left\{\frac{K\left(\frac{t-s}{b}\right)}{Y(s)}\right\}^2\mathrm{d}N(s)\\ &= b^{-2}\sum_i K\left(\frac{t-T_i}{b}\right)^2\frac{\Delta N(T_i)}{Y(T_i)^2}\end{aligned} \tag{3.14}$$

Two typical kernel functions include the Epanechnikov kernel

$$K_E(x) = \frac{3}{4}(1-x^2), \qquad -1 \leq x \leq 1$$

and the biweight kernel

$$K_B(x) = \frac{15}{16}(1-x^2)^2, \qquad -1 \leq x \leq 1.$$

By assuming an Epanechnikov kernel and observing that $K_E(x)^2 = 3/5 \times K_B(x)$, Equation (3.14) becomes

$$\begin{aligned}\sigma^2[\hat{\alpha}(t)] &= b^{-2}\sum_i K_E\left(\frac{t-T_i}{b}\right)^2\frac{\Delta N(T_i)}{Y(T_i)^2}\\ &= \frac{3}{5}b^{-1}\left[b^{-1}\sum_i K_B\left(\frac{t-T_i}{b}\right)\frac{\Delta N(T_i)}{Y(T_i)^2}\right]\end{aligned}$$

so that the variance of $\hat{\alpha}$ can be estimated from a (standard) kernel smoothing of $\Delta N(T_i)/Y(T_i)^2$ using a biweight kernel, with an adjustment by a scale factor of $3/(5b)$. This approach has not been made explicit in standard texts on survival analysis (e.g. Klein and Moeschberger, 1997) or density estimation (e.g. Wand and Jones, 1995).

3.6.2.4 Density and hazard estimation: a link

There is a useful connection between weighted density estimation and hazard estimation. For the unweighted case, the density function f for observations T_i is estimated by

$$\hat{f}(t) = b^{-1} \sum_i K\left(\frac{t - T_i}{b}\right)$$

and, for the weighted case, with weights w_i, the density function is estimated by

$$\hat{f}(t) = \frac{1}{b \sum w_i} \sum_i K\left(\frac{t - T_i}{b}\right) w_i$$

By letting the weights in the weighted density estimation be the steps in the cumulative intensity, the hazard can be estimated from the weighted density estimator multiplied by the sum of the steps.

This close relationship allows methodological and theoretical developments in density estimation to be applied to hazard estimation (see Wand and Jones, 1995).

As a novel application of this relationship, we can use `locfit` for weighted density estimation to estimate the intensity function and its variance for left truncated and right censored weighted data.

3.6.2.5 Intensity estimation over the Lexis diagram

Smoothing by age and by cohort, or by period, follows McKeague and Utikal (1990) and Keiding (1990). The main connection is that hazard regression with a covariate is equivalent to smoothing over the surface defined by the primary time variable and the covariate. Keiding suggested performing two separate kernel smoothings, each on two-dimensions. First, the risk set $Y(a,t)$ is smoothed and then $\mathrm{d}A = \Delta N/Y$ is smoothed, using the smoothed Y.

A novel approach, proposed to deal with the issue of the boundary, is to transform from the unit triangle (bounded by (0,0), (1,0) and (0,1)) to the unit square. Moreover, it is proposed to use software for local likelihood density estimation, allowing for weights, to estimate the smoothed Y and $\Delta N/Y$.

Specifically, the smoothing of Y follows McKeague and Utikal (1990), where the birth cohorts (in this context, the covariate) are partitioned into small bands. Then $Y(a,t)$ is calculated for the mid-point of the Lexis rhombus with corners defined by

$Y(a-\delta/2,t)$, $Y(a+\delta/2,t)$, $Y(a+\delta/2,t+\delta)$ and $Y(a+3\delta/2,t+\delta)$ with interval width δ.

The smoothing parameter in the `locfit()` function was estimated using generalised cross-validation as described in the monograph by Loader (1999).

To estimate the hazard function over the Lexis diagram using retrospective data, the age-cohort data were transformed to the unit square for estimation and then transformed back. This is a multivariate version of the transformation kernel density estimators discussed by Wand and Jones (1995). The transformation g used was

$$g(a,c) = (a + \min(a-a_0, c-c_0),\ c + \min(a-a_0, c-c_0)).$$

The Jacobian of this transformation is 2 (Mood et al., 1974). The hazard function α can then be represented by the Jacobian times the hazard function on the transformed data (λ_g):

$$\alpha(a,t) = \alpha(a, a+c) = 2\lambda_g(g(a,c)).$$

When both the risk set and steps are smoothed across the Lexis diagram, no adjustment is required as the Jacobian is in both the numerator and the denominator.

This method was validated using a simulation dataset (see Section 5.A on page 155). An alternative method is to stratify on 10-year birth cohorts, perform separate analyses, and then use mid-cohort estimates to derive the surface on the Lexis diagram. This alternative method, although useful for validation as it is less sensitive at the boundaries, was considered an *ad hoc* solution and not followed further.

3.6.2.6 Semi-parametric estimation with covariates

A useful alternative approach is to use the well known proportional hazards model (Therneau and Grambsch, 2000). This semi-parametric model assumes a non-parametric baseline hazard function $\alpha_0(t)$ and proportionality between groups defined by covariates $\boldsymbol{Z}$ (e.g. birth cohort), such that

$$\alpha(t;\, \boldsymbol{Z}) = \alpha_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{Z}).$$

The model provides a convenient regression framework for model development. For assessment of the functional form for the covariates, penalised splines can be used (Therneau and Grambsch, 2000). Care is required to assess whether the proportionality assumption has been violated. Non-proportionality for smoking initiation or cessation can be interpreted as age-specific smoking transitions for a period being different from earlier periods. A good example is smoking uptake by older women in the middle half of the twentieth century (see Chapter 5). The use of proportional hazards will be investigated for estimation of cessation and initiation rates due to the convenience it would offer for extrapolation to earlier or later cohorts.

3.6.2.7 Bias in retrospective prevalence estimates

To investigate the bias in retrospective data it is useful to consider former and current smokers together as ever smokers (represented by E and subscript e). Assume observation at time t and estimates required for a fixed cohort c at an earlier time s. Let the prevalence of ever smoking at time t be represented by $\pi_e(t)$, where the explicit dependence upon age has been dropped for notational convenience. Let the observed prevalence conditional upon survival to time t of current and ever smoking at time s be $\pi_c^*(s)$ and $\pi_e^*(s)$, respectively. Let the transition probability $P_{ab}(s,t)$ be the probability of moving from state A at time s to state B at time t, and let the transition probability $P_{ab}^*(s,t)$ be defined as the probability of moving from state A at time s to state B at time t conditional upon being observed at time t. Let $S_y(s,t)$ be the probability of survival from time s to time t given in state Y at time s.

In the following development, the dependence of the transition probabilities and survival on s and t has been dropped for notational convenience. The observed prevalence of ever smoking at time s has a simple relationship with the transition probabilities, where

$$\pi_e^*(s) = \frac{\pi_e(s)P_{ee}}{\pi_e(s)(P_{ee}+P_{en}) + [1-\pi_e(s)]\cdot(P_{nn}+P_{ne})}.$$

which accounts for differential mortality and smoker recall error. Rearrangement of this equation provides a mechanism to adjust the retrospective estimate for bias, where

$$\pi_e(s) = \frac{\pi_e^*(s)/S_e}{\pi_e^*(s)/S_e + [1-\pi_e^*(s)]/S_n - P_{en}/S_n/S_e}.$$

For current smokers, a similar relationship is found, where

$$\pi_c(s) = \frac{\pi_c^*(s)/S_c}{\pi_c^*(s)/S_c + [1-\pi_c^*(s)]/(S_x+S_n) - P_{cn}/S_c/(S_x+S_n)}. \tag{3.15}$$

Equation (3.15) can then be interpreted as a modification of the classic equation due to Harris (1983, Equation 1), with an appropriate adjustment for former smoker recall error. Without the adjustment for recall error, the estimated prevalence of ever smokers would be biased downward.

By specification of the transition probabilities and the observed retrospective prevalence, the actual prevalence can be calculated. Multi-state models are one approach to calculate the transition probabilities, allowing for initiation, cessation and recall error. However a variety of other methods have been suggested to calculate the adjustment (e.g. Harris, 1983; Brenner, 1993; Burns et al., 1997b).

It is useful to relate prevalence at time t with the observed prevalence at time s, where

$$\pi_e(t) = P^*_{ee}\pi^*_e(s) + P^*_{ne}[1 - \pi^*_e(s)] \tag{3.16}$$

Burns et al. (1997b) used a novel method to estimate differential mortality. The authors assumed that the only significant influence on ever smoking for those aged 30 years and over was differential mortality. By fitting an appropriate curve to current status data for a cohort, a correction factor $I(s, t)$ for the apparent effect due to differential mortality was calculated using $I(s, t) = \pi_e(s)/\pi_e(t)$. Let the authors' corrected prevalence be $\tilde{\pi}_e(s)$. Using Equation (3.16), we have

$$\begin{aligned}\tilde{\pi}_e(s) &= \pi^*_e(s)\, I(s, t) \\ &= \pi^*_e(s)\frac{\pi_e(s)}{\pi_e(t)} \\ &= \pi_e(s)\frac{1 - P^*_{ne}/\pi_e(t)}{1 - P^*_{en} - P^*_{ne}}\end{aligned}$$

which will be a biased estimator of the actual prevalence unless initiation and recall error are either negligible or they cancel out each other (which is satisfied when $\pi_e(t) = P^*_{ne}/(P^*_{en} + P^*_{ne})$). In practice, initiation after age 30 years for males was relatively low, however initiation for older female cohorts after that age was not negligible. As will be seen from Chapter 6, recall error by former smokers could be relatively large. The transition probabilities for the survivors may be different from those for the cohort, as they may be affected by differential mortality. The likely effect of any bias would be to increase the probability of recall error, as those former smokers who are identified as being never smokers are expected to be more likely to survive.

One potential concern with the underlying model is the possibility for "loop feedback", where former smokers who have changed their recall to being never smokers are potential candidates for uptake. The effect of this is likely to be small, as there will be few who will change their recall early enough to take up, given the slow (assumed) rate of misclassification and the restricted uptake after age 30 years. Alternatively, smoking experimentation in earlier years could be restricted more strongly.

In summary, retrospective prevalence estimates are potentially biased due to differential mortality and recall error. Unfortunately, previous efforts to adjust retrospective estimates have tended to focus solely upon differential mortality. In particular, the method for adjustment proposed by the 1997 National Cancer Institute smoking monograph may be biased (Burns et al., 1997b). A multi-state modelling approach will be used in later chapters rather than use the adjustments given in this section.

3.6.2.8 Bias in retrospective hazard estimates

It is well recognised that retrospective hazard estimates with differential mortality are biased (Keiding, 1991, Section 6.2). As an intuitive motivation for this relationship, consider an informal definition for the transition intensity, where

$$\alpha_{jk}(s) = \frac{\mathrm{d}N_{jk}(s)}{Y_j(s-)}.$$

Under differential survival, those individuals alive at time $s-$ will have a different probability of survival compared with those who change from state j to state k in $[s, s + \mathrm{d}s)$. The transition intensity, conditional upon survival to a later time t, can then be described (informally) by

$$\begin{aligned} & \alpha_{jk}(s \mid \text{survival to time } t) \\ = \; & \frac{\mathrm{d}N_{jk}(s \mid \text{survival to time } t)}{Y_j(s- \mid \text{survival to time } t)} \\ = \; & \alpha_{jk}(s)\frac{P(\text{Survival to time } t \mid \text{State } k \text{ at time } s)}{P(\text{Survival to time } t \mid \text{State } j \text{ at time } s-)}. \end{aligned} \tag{3.17}$$

However this development does not include any effect due to misclassification. Using the same notation as the previous section, and letting the observed rates of initiation and cessation at time s conditional upon observation at time t be $\alpha_I^*(s)$ and $\alpha_Q^*(s)$, respectively, we find that

$$\begin{aligned}
\alpha_I^*(s) &= \alpha_I(s)\frac{S_c - P_{cn}}{S_n} \\
\alpha_Q^*(s) &= \alpha_Q(s)\frac{S_x - P_{xn}}{S_c - P_{cn}}.
\end{aligned}$$

The adjustments in Equation (3.17) are used in the analysis of retrospective data in Chapter 5, as estimates for misclassification rates were not available until Chapter 6.

3.6.3 Current status estimation

Current status data are based on the current state for a given factor. For smoking, an individual's current state could be a never, current or former smoker. Data from a single cross-sectional survey can be analysed by making stationarity assumptions. This approach has been reviewed by Diamond and McDonald (1991), where their emphasis was on covariates rather than the intensity itself. For estimation of the intensity function, see Keiding (1991). The stationarity assumption assumes that age-specific changes in smoking are stable over time, which in practice is often not true. As a consequence, current status data from multiple points in time are required for valid estimation.

Current status data are not as efficient as retrospective data because all current status data are censored: the event either happened prior to the given time (right censored) or the event has not happened (left censored).

The main advantage of current status data is improved validity compared with retrospective data, which is argued convincingly by Diamond and McDonald (1991).

The analysis of current status data followed here borrows heavily from previous efforts to analyse serial prevalence studies (Ades and Nokes, 1993; Marschner, 1997). The general approach is to use a binomial likelihood for the observed prevalence and to model the prevalence as a function of the exponential of the cumulative hazard. A variety of forms for the hazard have been proposed. Extensions include the use of the EM algorithm to model for differential selection (Marschner, 1997), such as due to differential mortality.

In the following, current status data are available from one or more cross-sectional surveys. Typically, the data are available by age and sex.

One approach is to consider the change in prevalence, for which there are some relatively simple relationships. If $T = N + C + X$ is the total live population with no migration, then the mortality rate μ can be represented by

$$\frac{\partial T}{\partial t} + \frac{\partial T}{\partial a} = -\mu T \tag{3.18}$$

Then from the differential equations in Equations (3.6)–(3.8) and using the relationships that $\pi_n = N/T$, $\pi_c = C/T$ and $\pi_x = X/T$, we have that

$$\begin{aligned}
\frac{\partial \pi_n}{\partial t} + \frac{\partial \pi_n}{\partial a} &= \frac{N'T - NT'}{T^2} \\
&= -\left[\mu_0 + \alpha_I - \mu\right]\pi_n + \alpha_E \pi_x && (3.19) \\
\frac{\partial \pi_c}{\partial t} + \frac{\partial \pi_c}{\partial a} &= -\left[RR_c\mu_0 + \alpha_Q - \mu\right]\pi_c + \alpha_I \pi_n && (3.20) \\
\frac{\partial \pi_x}{\partial t} + \frac{\partial \pi_x}{\partial a} &= -\left[RR_x\mu_0 + \alpha_E - \mu\right]\pi_x + \alpha_Q \pi_c && (3.21)
\end{aligned}$$

This gives the simple relationship that the smoking cessation rate can be estimated from rate of change of prevalence of current smoking together with the mortality rate difference between current smokers and the total population, with a small adjustment for smoking uptake, where

$$\alpha_Q = -\frac{\mathrm{d}\pi_c}{\mathrm{d}t}/\pi_c - (RR_c\mu_0 - \mu) + \alpha_I \frac{\pi_n}{\pi_c}. \tag{3.22}$$

In practice, the adjustment for smoking uptake at older ages may be small or an estimate will be required from an additional data source.

A similar relationship for smoking initiation rates is useful if an exogenous estimate of the recall error is available:

$$\alpha_I = -\frac{\mathrm{d}\pi_n}{\mathrm{d}t}/\pi_n - (\mu_0 - \mu) + \alpha_E \frac{\pi_x}{\pi_n}. \tag{3.23}$$

A similar relationship for former smokers can also be derived, however in practice it is less useful, because changes in uptake and misclassification make estimation difficult. This issue will be considered in more depth in Chapter 7.

These associations can be interpreted by the differential equations together with an adjustment for the change in total size of the population.

3.6.3.1 Rate of decline of prevalence

Let prevalence for a given state be represented by $\pi(a,t)$ and let the cohort-specific rate of decline of prevalence be represented by γ where

$$\gamma = -\frac{\mathrm{d}\pi}{\mathrm{d}t}/\pi = -\left[\frac{\partial\pi}{\partial a} + \frac{\partial\pi}{\partial t}\right]/\pi. \tag{3.24}$$

For cohort-specific prevalence between time s and time t, the estimate of the mean rate of decline of prevalence $\bar{\gamma}$ is

$$\bar{\gamma} = \frac{1}{t-s}\log\frac{\pi(a,s)}{\pi(a+t-s,t)}.$$

If prevalence is observed for a cohort at time s and time t with effective cell sizes (Rao and Scott, 1992) of $n(a,s)$ and $n(a+t-s,t)$, respectively, the standard error for $\bar{\gamma}$ is estimated (Rothman and Greenland, 1998) by

$$se(\hat{\bar{\gamma}}) = \frac{1}{t-s}\sqrt{\frac{1-\pi(a,s)}{n(a,s)\pi(a,s)} + \frac{1-\pi(a+t-s,t)}{n(a+t-s,t)\pi(a+t-s,t)}}.$$

Note that these would in practice be estimates for aggregate age groups or cohorts, which would imply a step function for age.

This simple approach can be used with large cell sizes such as provided by Census data, however small cell sizes such as those typically found from using survey data require the use of various regression approaches. For ease of variance estimation, it is advantageous to use a regression model with a log link, such that

$$\pi = \exp(\boldsymbol{\beta}'\boldsymbol{x})$$

where $\boldsymbol{\beta}$ is a vector of parameters and $\boldsymbol{x}$ is a vector of covariates. The change in prevalence hazard can be estimated from the prevalence data and expressed by the regression coefficients, where

$$\gamma = -\frac{\mathrm{d}\pi}{\mathrm{d}t}/\pi = -\boldsymbol{\beta}'\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}. \tag{3.25}$$

Typically, the model can be fitted using standard generalised linear regression routines, using weighted data, with binomial errors and a log link. Standard errors can be estimated by

$$se(\hat{\gamma}) = \sqrt{\boldsymbol{u}'\boldsymbol{\Sigma}_{\hat{\beta}}\boldsymbol{u}}, \tag{3.26}$$

where $\boldsymbol{u} = -\mathrm{d}\boldsymbol{x}/\mathrm{d}t$ and $\boldsymbol{\Sigma}_{\hat{\beta}}$ is the covariance matrix for $\hat{\boldsymbol{\beta}}$. Issues with log-binomial modelling are briefly reviewed in the following section on page 82.

At present, local likelihood estimation for binomial data using the `locfit` library only allows a logit link (Loader, 1999). Following a similar development, this is a logistic model

$$\pi = \text{expit}(\boldsymbol{\beta}'\boldsymbol{x})$$

where again $\boldsymbol{\beta}$ is a vector of parameters and $\boldsymbol{x}$ is a vector of covariates, and where $\text{expit}(y) = 1/[1-\exp(-y)]$. The change in prevalence hazard can be estimated from the prevalence data and expressed by the regression coefficients, where

$$\gamma = -\frac{\mathrm{d}\pi}{\mathrm{d}t}/\pi = -(1-\pi)\frac{\mathrm{d}\boldsymbol{\beta}'\boldsymbol{x}}{\mathrm{d}t}. \tag{3.27}$$

One advantage of using `locfit` is that the derivatives of the linear predictor can be estimated directly using the second parameter of the local polynomial (see Loader, 1999). Estimation of standard errors using a logit link is more complicated than using a log-binomial model. One approach for variance estimation would be to use the bootstrap (Davison and Hinkley, 1997).

3.6.3.2 Implications of log-binomial modelling

Log-binomial modelling of prevalence provides a simple method to estimate variance for the rate of change of slope in the previous section. However there has been some debate in the literature about appropriate modelling of prevalence to estimate prevalence proportion ratios. A recent comparison of methods used simulations to show that the log-binomial regression provided unbiased point estimates with correct type I error probabilities (Skov et al., 1998).

The main criticism of log-binomial regression is that the linear predictor must be less than or equal to zero ($\beta' x \leq 0$) to ensure the predicted prevalence takes values equal to or less than one ($\pi = \exp(\beta' x) \leq 1$). Skov et al. (1998) used a geometry argument to show that interpolation of covariates provides predicted values less than one. Extrapolation of covariates remains an issue.

A second criticism, relating to the use of a non-canonical link function, is that the standard errors of the parameter estimates may be biased in an unknown direction (Ma and Wong (1999) citing the monograph by Collett (1991)). One possible explanation for this criticism is that the confidence intervals should only include predicted values less than one. One practical approach is to interpret with care any interval that lies above one. Alternatively, bootstrapping would be used for confidence interval estimation given a constrained interval.

An alternative method is to use the complementary log-log link (Ma and Wong, 1999), which was not considered by Skov et al. (1998). However, the method models the rate ratios, and estimation of both the rates and the differentials of the rates, as required in this thesis, would be difficult.

For modelling cross-sectional data, a synthesis of the literature suggests: prevalence proportions are best estimated using logistic regression; hazard ratios from disease prevalence can be best estimated using generalised linear models with binomial errors and the complementary log-log link; and the choice of method to estimate hazards and prevalence proportion ratios is equivocal.

3.6.4 Fitting dynamic models

For census data, it is possible to fit a dynamic model using the prevalence of both current and former smokers. The model fitting can be performed using maximum likelihood and a Newton-like procedure to estimate the parameters. The likelihood function specifies the probability distribution for the prevalence data that is fitted to the full dynamic model using a set of parameters ($\boldsymbol{\theta}$).

Let the observed data for time s to time t be

$$\begin{aligned}\pi_c &= y_c/n \\ \pi_x &= y_x/n\end{aligned}$$

where n is the total number observed, y_c and y_x are the number of current and former smokers, respectively, and π_c and π_x are the prevalence of current and former smokers, respectively.

Then the model can be represented by

$$\begin{aligned}\pi_c(s;\boldsymbol{\theta}) &= \operatorname{logit}(\beta_c) \\ \pi_x(s;\boldsymbol{\theta}) &= \operatorname{logit}(\beta_x) \\ \pi_c(t;\boldsymbol{\theta}) &= \sum_j P(\text{State } C \text{ at time } t \mid \text{State } j \text{ at time } s \text{ and survived}; \boldsymbol{\theta}) \\ &\qquad \times P(\text{State } j \text{ at time } s; \boldsymbol{\theta}) \\ \pi_x(t;\boldsymbol{\theta}) &= \sum_j P(\text{State } X \text{ at time } t \mid \text{State } j \text{ at time } s \text{ and survived}; \boldsymbol{\theta}) \\ &\qquad \times P(\text{State } j \text{ at time } s; \boldsymbol{\theta})\end{aligned}$$

where the log-likelihood, assuming a multinomial distribution for the observed prevalence of current and former smoking, is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = \sum_{\text{time } \in\{s,t\}} & \left\{ \log\left[\frac{n!}{y_c! y_x! (n-y_c-y_x)!}\right]\right. \\ & + y_c \log(\pi_c) + y_x \log(\pi_x) \\ & \left. + (n-y_c-y_x)\log(1-\pi_c-\pi_x) \right\}.\end{aligned}$$

As indicated earlier, this likelihood is maximised using a Newton-like algorithm.

3.6.5 Survey data

Most of the information on smoking comes from sample surveys that have complex designs. Valid analysis of these data must take some account of the sample weights and the design of the surveys, including stratification and primary sampling unit information (Cochran, 1977).

Survey analysis software, such as `SUDAAN` and procedures in `Stata` and `SAS`, were available for some of the estimation procedures. However the use of these tools were constrained for three reasons. First, there were limitations on software available to perform the proposed analyses for only weighted data, ignoring any clustering effects. Second, weighting and design information were not available for some of the data sources, including the ACCV Australian prevalence data (see Chapter 2). Third, the study designs varied between data sources, making comparisons difficult. In particular, the Census data were treated as being independently sampled, which did not fit well within a survey analytical framework.

When survey design information and suitable software were available, that analysis was undertaken. In other situations, a pragmatic decision was made to perform an independent weighted analysis assuming a conservative design effect. The adjusted weights w_i^*, which are equivalent to effective cell sizes (Rao and Scott, 1992), were estimated by

$$w_i^* = \frac{n}{D_{\text{eff}}} \cdot \frac{w_i}{\sum_{j=1}^{n} w_j}$$

where n is the number of observations, w_i are the sample weights and D_{eff} is the approximate design effect. Obvious deficiencies include that the design effect may not be appropriate for a given survey, for a particular stratum, or for a particular outcome variable.

3.6.6 Matrix exponentiation

For well-conditioned matrices, matrix exponentiation can be found directly using spectral decomposition. However this is sensitive to ill-conditioned matrices (Stewart, 1994). The approach taken here is to use numerical routines for solving ordinary differential equations to estimate the matrix exponential. This is particularly useful, as the transition matrix does not have to be expressed in its analytical form. This allows for more general models and avoids a complicated and potentially intractable analytical derivation.

The Fortran code used was `VODE`, accessed from `Netlib`, which uses backward differentiation formulae due to Brown et al. (1989). Wrapper subroutines were written to make the code accessible in `R` statistical software. Compilation was carried out using `g77`, which is part of the GNU Compiler Collection. In practice, the compilation was carried on the Windows 2000 operating system, however all of the software used was open source and would compile similarly on Unix or Linux operating systems.

Results were validated against estimates using spectral decomposition for well-conditioned matrices. The development of this tool provided a robust method for matrix exponentiation for matrices up to dimensions of 50 × 50.

Example `R` code for matrix exponentiation is (`#` denotes a comment, `>` denotes the command prompt):

```
> require(linode)  # use the linode library
> intensity.mat <- matrix( c(-0.01,  0.01, 0,
                             0,    -0.03, 0.03,
                             0.005, 0,   -0.005 ),
                           byrow=T, nrow=3)  # define the intensity matrix
> mexp(intensity.mat)  # matrix exponential of the intensity matrix
             [,1]         [,2]         [,3]
[1,] 9.900501e-01 9.802205e-03 0.0001477442
[2,] 7.384536e-05 9.704457e-01 0.0294803948
[3,] 4.962650e-03 2.461875e-05 0.9950127288
```

Recently the `R` package `odesolve` was released for solving ordinary differential equations. This is substantially slower than the solution proposed here (for 100 matrix exponentials of 3 × 3 matrices, `odesolve` took 36s compared with 6s for `linode`). However the `odesolve` package is quite general and could deal with matrices of 300 × 300 or larger.

3.6.7 Local likelihood

Local likelihood is a recent analytical method that is used throughout this thesis. It is used variously for modelling hazard densities, prevalence and mortality rates. The following discussion is strongly influenced by Loader (1999).

As an outline of the method, the likelihood function is defined *locally* to a fitted point x. The fitted parameters a minimise the local log-likelihood $\mathcal{L}_x$ for x, where

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x) l(Y_i, \theta(x_i))$$

where l is the local log-likelihood component for observation i, w is a weight function defined on a bandwidth with an associated weight function, and $\theta(x_i)$ is defined as

$$\theta(x_i) = a_0 + a_1(x_i - x) + \frac{1}{2}a_2(x_i - x)^2.$$

The approach can be described as taking a local approach to likelihood estimation (hence the name), where the likelihood at a point is constructed by weighting data points close to the fitted point.

An important aspect of model selection is the smoothing parameter that determines the bandwidth, which can be interpreted in terms of asymptotic degrees of freedom for the model. Loader (1999) recommends the use of generalised cross validation.

The main theoretical advantage for local likelihood over related kernel approaches is the existence of small sample properties. Another advantage is that good software has been implemented that allows practical use of the theory. The software is available as the `locfit` package in `R`.

For Chapter 1, local likelihood estimation was used to smooth the age-specific rates and prevalence. The prevalence modelling included binomial errors and a logit link. The rate modelling included Poisson errors and used a log link. The rate ratios in Figure 1.14 were estimated using rates as outcomes, with separate analyses by sex and study. Covariates in the rate ratio model included age and an interaction between age and current smoking. The rate ratios were estimated using the log of the interaction terms.

3.6.8 Commentary on quit ratios as a measure of quitting

The most common measure of cessation is the *quit ratio*, defined as the proportion of ever smokers who are ex-smokers (Pierce et al., 1987). The quit ratio for quitting activity is analogous to smoking prevalence for smoking activity (U.S. Department of Health and I

1989), where both measures have as their denominator those eligible for the given activity.

The quit ratio is technically not a ratio, where former smokers are included in both the numerator and the denominator (U.S. Department of Health and Human Services, 1990). Although the term remains in common use, a better name may be the *prevalence of cessation* (Giovino et al., 1995).

The measure is simple and easily estimated from cross-sectional data, however there are limitations on its interpretation. Ex-smokers are treated as being equal, irrespective of whether some may have quit recently and may restart. Similarly, current smokers are treated equally, although some may have restarted after a period of cessation. Moreover, the quit ratio is a cross-sectional summary of a dynamic process, and does not necessarily express recent changes.

Comparisons between quit ratios can be difficult to interpret. For instance, as a population ages then there is more opportunity to begin to smoke, to quit smoking, and for differential mortality due to smoking. There is greater ease in comparing groups that are alike, such as trends for an age group.

To provide some technical detail, let $QR(t)$ represent the quit rate for time t. The quit ratio is defined as

$$QR = \frac{\pi_x}{\pi_c + \pi_x},$$

and the change in the quit ratio, expressed as the derivative with respect to time t, is

$$\frac{\mathrm{d}QR}{\mathrm{d}t} = \frac{\pi_x' \pi_c - \pi_x \pi_c'}{(\pi_c + \pi_x)^2}.$$

A necessary and sufficient condition for no change in the quit ratio is

$$\frac{\pi_c'}{\pi_c} = \frac{\pi_x'}{\pi_x},$$

such that the proportional change in current smokers is the same as for ex-smokers.

For a varying proportional change for current and ex-smokers, consider the odds ratio of the quit ratio at time 0 and time t:

$$\text{Odds ratio of } QR = \frac{QR(t)}{1 - QR(t)} \cdot \frac{1 - QR(0)}{QR(0)} = \frac{\pi_c(0)}{\pi_c(t)} \cdot \frac{\pi_x(t)}{\pi_x(0)}.$$

This can be fitted using logistic regression for the quit ratio against t. Other covariates, possibly with interactions, can also be included in the regression. The odds ratio is estimated using the regression coefficient (β) for t by $\exp(\beta t)$, giving the relationship:

$$\frac{\pi_c(t)}{\pi_c(0)} = \frac{\pi_x(t)}{\pi_x(0)} \exp(-\beta t).$$

Therefore, the proportional change in current smoking prevalence equals the proportional change in ex-smoking prevalence times the inverse of the odds ratio of the quit ratios.

3.6.8.1 Process interpretation

It would be useful to have a process interpretation for changes in the quit ratios. For comparing two similar populations over time, let it be assumed that smoking uptake and mortality for never, current and ex-smokers change slowly over time. Changes in smoking uptake can be assumed negligible for populations aged 25 years and over. The assumption of constant mortality rates may not hold for older age groups. Let ϕ be the *change* in instantaneous quit rate between the two populations. These assumptions can be expressed as

$$\begin{aligned}\frac{\mathrm{d}C}{\mathrm{d}t} &= -\phi C\\ \frac{\mathrm{d}X}{\mathrm{d}t} &= \phi C\end{aligned}$$

so that the change in one minus the quit ratio is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(1-QR) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{C}{C+X}\right]\\ &= \frac{C'X - X'C}{(C+X)^2}\\ &= -\phi\frac{C}{C+X}\\ &= -\phi(1-QR).\end{aligned}$$

This ordinary differential equation can be expressed as

$$\frac{C}{C+X} = \frac{C(0)}{C(0)+X(0)}\exp(-\phi t),$$

which can be fitted as a log-binomial regression model, with the coefficient for t being an estimate of the change of the quit rate. A simple estimate of ϕ is

$$\phi = \frac{1}{t} \log \frac{1 - QR(0)}{1 - QR(t)}.$$

Given that ϕ will often be small,

$$\phi \approx \frac{QR(t) - QR(0)}{t},$$

so that the slope of a linear regression of the quit ratios is approximately the change in the instantaneous quit rate between the two populations.

A similar argument can be followed for smoking prevalence, with the additional assumption that $dN/dt = 0$, so that the slope for changing prevalence gives an estimate of the change in quit rate. The use of the quit ratios avoids having to make assumptions about never smokers.

The assumption of constant mortality may not hold. In that case, the arguments here still hold if $-C'/C \approx N'/N$, whereupon ϕ can be interpreted as the change in the rate of decline of current smokers. Alternatively, the condition that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[\frac{C}{C+X}\right] = -\phi\frac{C}{C+X}$$

will allow the arguments to hold, with ϕ now being interpreted as the change in the rate of decline of the prevalence of current smokers.

The quit ratio, or prevalence of cessation, is an analogue to smoking prevalence for quitting behaviour. Similarly to smoking prevalence, the quit ratio is a cross-sectional summary of a dynamic process, and comparisons between populations are difficult to interpret, particularly between age groups. Changes over short periods of time for an age-specific group approximate the change in the instantaneous quit rate.

3.6.9 Misclassification of former smokers

For a worst-case scenario, consider all cause mortality by the usual equation

$$\mu = \mu_0 \left[\pi_c RR_c + \pi_x RR_x + \pi_n\right].$$

Now, assume a proportion α of the former smokers are misclassified as never smokers. As a worst case, assume that those former smokers who are misclassified have the average mortality rate of all former smokers. Algebraically, this is

$$\mu = \mu_0 \left[\pi_c RR_c + (1-\alpha)\pi_x RR_x + (\pi_n + \alpha\pi_x RR_x)\right].$$

However, the observed mortality rate for "never" smokers is now $\mu_0(\pi_n+\alpha\pi_x RR_x)/(\pi_n+\alpha\pi_x)$ and the actual rate ratios for RR_x and RR_c are scaled by a factor

$$\frac{\pi_n+\alpha\pi_x}{\pi_n+\alpha\pi_x RR_x}=\frac{1}{1+\alpha\frac{\pi_x}{\pi_n}(RR_x-1)}.$$

For a more extreme situation in which $RR_x \approx 1.5$, $\alpha \approx 0.1$ (10% misclassification) and $\pi_x = 2\times\pi_n$, the worst case scenario gives a scale factor of 0.91, or a 9% reduction in the actual rate ratios for current and former smokers. More realistic values of α may be 0.05 and $\pi_x \approx \pi_n$ giving a scale factor of 0.975, or a 2.5% reduction in the rate ratios. However this still assumes that the misclassified former smokers were at comparable risk to the average former smoker. This may be quite unlikely, as the misclassification may be due to long-term cessation, where the person no longer identifies themselves as being a former smoker.

In conclusion, rate ratios are biased due to smoker misclassification, however in practice the level of bias is acceptable for population modelling. Bias due to misclassification is more of a problem for estimates of ever smoking prevalence (see Section 3.6.2.7 and Chapter 6).

3.A Analytical approximations for Model 1

As there is a cycle in Model 1 formed by initiation, cessation and recall error, it is difficult to represent the transition probabilities as integral equations. Analytical approximations can be obtained by ignoring the cycle in the model and assuming that only true never smokers can take up smoking. For P_{nn}, the time-heterogeneous case does include the first cycle for never smokers moving back to being never smokers again. In the following cohort analysis, for notational convenience only one time scale has been used.

$$
\begin{aligned}
P_{nn}(s,t) &= \exp\left[-\int_s^t (\alpha_I + \mu_0)\,\mathrm{d}u\right] \\
&\quad + \int_s^t \int_u^t \int_v^t \exp\left[-\int_s^u (\alpha_I + \mu_0)\,\mathrm{d}x\right] \alpha_I(u) \\
&\qquad \times \exp\left[-\int_u^v (\alpha_Q + RR_c\mu_0)\,\mathrm{d}x\right] \alpha_Q(v) \\
&\qquad \times \exp\left[-\int_v^w (\alpha_E + RR_x\mu_0)\,\mathrm{d}x\right] \alpha_E(w) \\
&\qquad \times \exp\left[-\int_w^t \mu_0\,\mathrm{d}x\right]\,\mathrm{d}w\,\mathrm{d}v\,\mathrm{d}u \\
P_{nc}(s,t) &= \int_s^t \exp\left[-\int_s^u (\alpha_I + \mu_0)\,\mathrm{d}v\right] \alpha_I(u) \exp\left[-\int_u^t (\alpha_Q + RR_c\mu_0)\,\mathrm{d}v\right] \mathrm{d}u \\
P_{nx}(s,t) &= \int_s^t \int_u^t \exp\left[-\int_s^u (\alpha_I + \mu_0)\,\mathrm{d}w\right] \alpha_I(u) \\
&\qquad \times \exp\left[-\int_u^v (\alpha_Q + RR_c\mu_0)\,\mathrm{d}w\right] \alpha_Q(v) \\
&\qquad \times \exp\left[-\int_v^t (\alpha_E + RR_x\mu_0)\,\mathrm{d}w\right]\,\mathrm{d}v\ \mathrm{d}u \\
P_{cc}(s,t) &= \exp\left[-\int_s^t (\alpha_Q + RR_c\mu_0)\,\mathrm{d}u\right] \\
P_{cx}(s,t) &= \int_s^t \exp\left[-\int_s^u (\alpha_Q + RR_c\mu_0)\,\mathrm{d}v\right] \alpha_Q(u) \exp\left[-\int_u^t (\alpha_E + RR_x\mu_0)\,\mathrm{d}v\right] \mathrm{d}u \\
P_{cn}(s,t) &= \int_s^t \int_u^t \exp\left[-\int_s^u (\alpha_Q + RR_c\mu_0)\,\mathrm{d}w\right] \alpha_Q(u) \\
&\qquad \times \exp\left[-\int_u^v (\alpha_E + RR_x\mu_0)\,\mathrm{d}w\right] \alpha_E(v) \\
&\qquad \times \exp\left[-\int_v^t \mu_0\,\mathrm{d}w\right]\,\mathrm{d}v\ \mathrm{d}u \\
P_{xx}(s,t) &= \exp\left[-\int_s^t (\alpha_E + RR_x\mu_0)\,\mathrm{d}u\right] \\
P_{xn}(s,t) &= \int_s^t \exp\left[-\int_s^u (\alpha_E + RR_x\mu_0)\,\mathrm{d}v\right] \alpha_E(u) \exp\left[-\int_u^t \mu_0\,\mathrm{d}v\right] \mathrm{d}u \\
P_{xc}(s,t) &= 0.
\end{aligned}
$$

For the time homogeneous case, the transition probabilities between time t and time $t+1$ $(= P_{jk}(t,t+1) \equiv P_{jk})$ are

$$
\begin{aligned}
P_{nn} &= \exp[-(\alpha_I + \mu_0)] + \frac{\alpha_I \alpha_Q \alpha_E}{3!} \\
P_{nc} &= \frac{\alpha_I}{\alpha_Q - \alpha_I + (RR_c - 1)\mu_0} \Big\{ \exp[-(\alpha_I + \mu_0)] - \exp[-(\alpha_Q + RR_c\mu_0)] \Big\} \\
P_{nx} &= \frac{\alpha_I \alpha_Q}{\alpha_E - \alpha_Q - (RR_c - RR_x)\mu_0} \\
&\quad \times \Bigg\{ \frac{1}{\alpha_Q - \alpha_I + (RR_c - 1)\mu_0} \Big[\exp[-(\alpha_I + \mu_0)] - \exp[-(\alpha_Q + RR_c\mu_0] \Big] - \\
&\qquad \frac{1}{\alpha_E - \alpha_I + (RR_x - 1)\mu_0} \Big[\exp[-(\alpha_I + \mu_0)] - \exp[-(\alpha_E + RR_x\mu_0] \Big] \Bigg\} \\
P_{cc} &= \exp[-(\alpha_Q + RR_c\mu_0)] \\
P_{cx} &= \frac{\alpha_Q}{\alpha_E - \alpha_Q + (RR_x - RR_c)\mu_0} \Big\{ \exp[-(\alpha_Q + RR_c\mu_0)] - \exp[-(\alpha_E + RR_x\mu_0)] \Big\} \\
P_{cn} &= \frac{\alpha_Q \alpha_E}{-\alpha_E - (RR_x - 1)\mu_0} \\
&\quad \times \Bigg\{ \frac{1}{\alpha_E - \alpha_Q + (RR_x - RR_c)\mu_0} \Big[\exp[-(\alpha_Q + RR_c\mu_0)] - \exp[-(\alpha_E + RR_x\mu_0] \Big] - \\
&\qquad \frac{1}{-\alpha_Q + (1 - RR_c)\mu_0} \Big[\exp[-(\alpha_Q + RR_c\mu_0)] - \exp[-\mu_0] \Big] \Bigg\} \\
P_{xx} &= \exp[-(\alpha_E + RR_x\mu_0)] \\
P_{xn} &= \frac{\alpha_E}{-\alpha_E + (1 - RR_x)\mu_0} \Big\{ \exp[-(\alpha_E + RR_x\mu_0)] - \exp[-\mu_0] \Big\} \\
P_{xc} &= 0.
\end{aligned}
$$

For the time homogeneous case, a short time period is assumed so that the cycle for never smokers back to never smokers has ignored mortality and used a result for birth processes.

Chapter 4

All cause mortality rate ratios for current and former smokers in Australia and New Zealand

Abstract

Aim: To describe the variation in all cause mortality rate ratios due to current and former smoking and estimate rate ratios that could be applied to smoking in Australia and New Zealand. **Methods:** A literature review and meta-regression of all cause rate ratios for current and former smokers was undertaken. Population-specific all cause mortality rate ratios were estimated from cause-specific mortality rate ratios combined with local smoking prevalence estimates and cause-specific mortality rates. **Results:** There was considerable variation in rate ratios by age, sex, over time and between source studies. There was evidence that estimates from the Cancer Prevention Study I and II may be unexpectedly high. Differences between Australian and New Zealand estimates based on the same set of cause-specific rate ratios and local prevalence and cause-specific rates were small. **Discussion:** Complete estimates for Australia and New Zealand are presented, but validity of the estimates is uncertain. Estimation of valid all cause mortality rate ratios for smoking in populations for which representative local data are not available is not a trivial task.

4.1 Introduction

The causal association between tobacco smoking and many diseases is well established (U.S. Department of Health and Human Services, 1989). Taking all of the associated diseases together, smoking exposure has a large adverse impact on population health.

All cause mortality rate ratios for current and former smokers are important variables in descriptive epidemiology. The rate ratios can be used to adjust for bias in retrospective estimates of smoking prevalence or smoking cessation rates data (see Harris, 1983; National Cancer Institute, 1997, and Section 3.6.2, pages 76–79). Moreover, multiple decrement life table analyses by smoking status (Rogers and Powell-Griner, 1991) and related dynamic models (Mendez et al., 1998) also require all cause mortality rates by smoking status.

In some situations the validity and precision of the rate ratio estimates are of less importance. For younger people, mortality rates are low so that differential mortality will have little impact on the estimates of the number or proportion by smoking status. However, with increasing age, cumulative smoking exposure together with higher mortality rates suggest that smoking estimates will be sensitive to estimates for the rate ratios.

It is well recognised that tobacco consumption patterns vary between populations (Todd, 1978) and within populations over time (Thun et al., 1997). Therefore it is not surprising that all cause mortality rate ratios vary by age, by sex and over time (Thun et al., 1997). Moreover, there have been changes over time in observed rate ratios for lung cancer and coronary heart disease (Thun et al., 1997), and cause-specific mortality rate ratios vary between populations (Peto et al., 1994; van de Mheen and Gunning-Schepers, 1996).

This suggests it may be difficult to estimate all cause mortality rate ratios for populations where no contemporaneous large cohort studies have been undertaken, such as in Australia and New Zealand.

My first objective is to assess the variability in reported all cause mortality rate ratios for current and former smokers compared with never smokers by sex, by age, and by time between populations. This will involve a meta-regression of the rate ratios, and in particular an assessment of whether there are any trends over time. This is similar in intent to a review by van de Mheen and Gunning-Schepers (1996) for variations in cause-specific mortality rate ratios.

In the presence of considerable variability between populations, my second objective is to investigate methods for population-specific rate ratio estimation. A method is described that uses cause-specific mortality rate ratios from large cohort studies together with population-specific, cause-specific rates and population-specific smoking prevalence when representative local data are not available. This method is then applied to Australian and New Zealand data.

4.2 Methods

4.2.1 Literature review

Two thorough reviews including meta-analyses of all cause mortality rate ratios for smoking were completed in 1990 (Holman et al., 1990) and 1995 (English et al., 1995). The two reviews were taken as providing adequate identification of relevant papers published before 1990.

To update the reviews, I conducted a Medline search, with publication year restricted to 1990–1999. Keywords include smoking or tobacco or cigarettes, deaths or mortality or all causes, and rates. The following sources were also investigated: descriptive studies that used all cause mortality rate ratios; publications from the larger cohort studies; and studies that estimated population attributable fractions. Referenced articles and citations in the Science Citation Index were searched for related studies.

The inclusion criteria for the review were for studies of Western populations that provided detailed information by sex and age group. The last criterion was to allow an investigation of changes in the rate ratios by age.

Estimation of variances for rate ratios followed methods used by English and colleagues (English et al., 1995). In outline, the variance for the log of rates was estimated from the width of the confidence interval, and the variance of the log of the rate ratio was estimated by the sum of the variances of the log of the rates. Some studies reported rates together with 95% confidence intervals that required aggregation. To do this, the number of deaths and the person-years of exposure were estimated from the rates and the confidence intervals.

4.2.2 Meta-regression

Results from the literature review were included in a meta-regression to describe the variation between the studies (Rothman and Greenland, 1998). Separate analyses were performed for males and females, with covariates for calendar period, age and some study indicators.

Weighted least-squares regressions were fitted using the log rate ratios as the dependent variable weighted by the inverse variances. Given the complexities of using weights in random effects models, all effects were treated as being fixed (Rothman and Greenlanc 1998). Separate analyses were performed for males and females, for current and for former smokers. There were too few studies for female former smokers and this combination was not modelled for covariates; only the average log rate ratio was estimated.

The base model used for all analyses included linear and quadratic effects for age, a linear term for time, indicator variables for the Cancer Prevention Study I (CPS-I) and the Cancer Prevention Study II (CPS-II), and an interaction term for linear effects between time and age. The indicators for CPS-I and CPS-II were included to assess whether these two important studies provided estimates that were appreciably different from those provided by other studies. The base model and all sub-models were fitted and compared for goodness of fit, choosing the model with the smallest Akaike's Information Criterion (Venables and Ripley, 1999).

The mid-point of the study period was taken as the value for time. For ages, the mid-point was again taken, excluding open age groups. Where multiple sets of results from the British Doctors Study were available, the results for current smokers for the periods 1951–1971 and 1971–1991 were included.

4.2.3 Estimation using cause-specific rate ratios

Precise cause-specific mortality rate ratios and confidence intervals were available from the CPS-II and from the combination of the 1986 National Mortality Followback Study and the 1987 National Health Interview Survey (NMFS/NHIS) (Malarcher et al., 2000). As a brief explanation of the design for the NMFS/NHIS, "case" exposure was taken from the National Mortality Followback Study, while the National Health Interview Survey provided population-level exposure. The causes of mortality of interest were lung cancer (ICD-9 code 162), coronary heart disease (ICD-9 codes 410–414), cerebrovascular disease (ICD-9 430–438), chronic obstructive pulmonary disease (ICD-9 490–492 and 496) and "other causes" (remainder of ICD-9).

Estimates for CPS-I, derived from tables of deaths and person-years of exposure by sex, age and cause (Burns et al., 1997c), are presented in Table 4.7 in Appendix 4.A.

Estimates for rate ratios for "other causes" for current smokers from CPS-II were estimated using rates and person-years used previously for indirect estimation of attributable mortality (Peto et al., 1992). Rate ratios for former smokers for "other causes" from CPS-II were estimated from similar estimates from CPS-I multiplied by the ratio of CPS-II to CPS-I estimates for "other causes" rate ratios for current smokers. The variance of the log of this rate ratio is the sum of the variance of the logs of the three components. These results are shown in Table 4.8 in Appendix 4.A. Rate ratios for "other causes" from NMFS/NHIS were assumed to be the same as for CPS-II. Estimates for cause-specific rate ratios are available from Malarcher et al.

(2000).

Population data and cause-specific mortality rates by year, age and sex for Australia and New Zealand were taken from the World Health Organisation Mortality Database (see Section 2.1.1 on page 29). The coding for chronic obstructive pulmonary disease from the database was for ICD-9 codes 490–496.

The theoretical development, including variance estimation, is given in the following section. In outline, all cause mortality rate ratios are estimated from the sum of cause-specific rate ratios weighted by the never smoker cause-specific mortality rates. Relative differences in the rate ratios between New Zealand, Australia and the contemporaneous Cancer Prevention Study were estimated separately for current and former smokers, by sex and year. Estimation was similar to that used for the formal meta-regression, with the inclusion of age as a factor and indicator variables for New Zealand and the relevant Cancer Prevention Study.

4.2.3.1 Technical development

An all cause mortality rate ratio for a population sub-group summarises a range of associations. A short development will show how these associations are related. To review notation, total mortality is represented by μ, current, former and never smoking prevalence are represented by π_c, π_x, π_n $(= 1 - \pi_c - \pi_x)$, and the mortality rate ratios for current and former smokers are given by RR_c and RR_x, respectively. The mortality rate for current, former and never smokers is given by $RR_c\mu_0$, $RR_x\mu_0$ and μ_0, respectively. There are also I groups of causes of death, indexed by i. Total mortality can be expressed as the sum of the mortality rates μ_i for each cause and divided by smoking status:

$$\begin{aligned}\mu = \sum_i \mu_i &= RR_c\mu_0\pi_c + RR_x\mu_0\pi_x + \mu_0\pi_n \\ &= \mu_0\left[(RR_c - 1)\pi_c + (RR_x - 1)\pi_x + 1\right]. \qquad (4.1)\end{aligned}$$

In turn, the mortality rates for each cause can be expressed by the rate ratios $RR_{c,i}$ and $RR_{x,i}$ and never smoker mortality rate $\mu_{0,i}$:

$$\begin{aligned}\mu_i &= \mu_{0,i}\left[(RR_{c,i} - 1)\pi_c + (RR_{x,i} - 1)\pi_x + 1\right] \\ \Leftrightarrow \mu_{0,i} &= \frac{\mu_i}{(RR_{c,i} - 1)\pi_c + (RR_{x,i} - 1)\pi_x + 1} \qquad (4.2)\end{aligned}$$

For causes that are not associated with tobacco smoking, $RR_{c,i} = RR_{x,i} = 1$ and $\mu_{0,i} = \mu_i$. By substitution and equating the components for each smoking status, we find the intuitive relationship that the mortality rate for never smokers is the sum of never smoker mortality rates for each cause. The rate ratios for current and ex-smokers also have simple expressions. By defining the weights $w_i = \mu_{0,i}/\sum_j \mu_{0,j}$ as the proportion of never smoker mortality accounted for by each cause, we have

$$\begin{aligned} \mu_0 &= \sum_i \mu_{0,i} && (4.3) \\ RR_c &= \sum_i w_i RR_{c,i} && (4.4) \\ RR_x &= \sum_i w_i RR_{x,i} && (4.5) \end{aligned}$$

Total mortality rate ratios for current and ex- smokers can be expressed in terms of observed cause-specific mortality rates and rate ratios together with prevalence. This can be done by substituting for $\mu_{0,i}$ and μ_0 from Equations (4.2) and (4.3) into Equations (4.4) and (4.5).

For variance estimation, we have the approximate relationship for RR (being either RR_c or RR_x):

$$\text{var}(RR) \approx \text{var}(\log RR) \times RR^2.$$

If we assume that the weighted sum of the cause-specific rate ratios are little affected by realistic changes in the weights, we find that

$$\text{var}\left(\sum_i w_i RR_i\right) \approx \sum_i w_i^2 \text{var}(\log RR_i) \times RR_i^2$$

for the rate ratio being either for former or for current smokers. By use of the Central limit theorem, the sum of variables will be approximately normally distributed, providing a mechanism for estimating the confidence interval. For tight confidence intervals of moderate rate ratios, the upper and lower bounds can be used to estimate the approximate variance of the log rate ratio.

4.2.4 Simple historical prevalence estimates

Estimates were required for 1966 and 1986, which are the central years for two published analyses of CPS-I and CPS-II, respectively.

Age-specific prevalence was estimated from the ACCV survey for Australia during 1986 (Hill, 1988). For New Zealand in 1986, and New Zealand and Australia in 1966, a combination of cross-sectional data and retrospective data were used to estimate prevalence. The data sources have been described in Chapter 2.

The approach was to use retrospective data to estimate the rate of change in age-specific prevalence, and then apply that change to the most recent year that complete age-specific prevalence data were available.

Naïve estimates of historical prevalence were required for all cause mortality rate ratio estimates based on cause-specific rates and rate ratios. Moreover, these simple estimates could be used as a cross-check of later estimates derived from fitting the multi-state model that had been driven backwards in time.

The approach was to take a reference year, for which complete data on age-specific prevalence were available, and the target year for which prevalence estimates were required. The change in age-specific prevalence from the reference year to the target was estimated and then applied to the age-specific prevalence available from the reference year.

The change function was estimated from retrospective data or current status data, dependent upon data availability and data quality. In general, prevalence was modelled over time on the logit scale using logistic regression, estimating the odds ratio for change in age-specific prevalence, where the odds ratio could possibly depend upon age. The basic relationship was

$$\log(\text{OR}) = \beta_0 \times \text{target} + \beta_1 \times \text{age} \times \text{target}$$

where "target" is a binary indicator for the target year. The model for prevalence used a factor for age in the logistic regression.

For the Australian data the prevalence for age groups 70–79 years and 80 years and over were estimated from an estimate for those aged 70 years and over. Data were taken from the 1976 New Zealand Census of Population and Dwellings to estimate the smaller age groups separately by sex. The scale factors to go from 70 years and over to 70–79 years and to 80 years and over were 1.07 and 0.74 for males, and 1.20 and 0.52 for females, respectively.

For Australia, the prevalence for 1966 was estimated using the prevalence for 1974 from ACCV data together with changes for 1966–1974 estimated using the National Heart Foundation Risk Factor Study.

For New Zealand, the reference years were 1976 and 1981, where Census data were available. The change functions were estimated using retrospective data for 1966–1976, and for 1981–1986, from the 1996/97 New Zealand Health Survey.

4.3 Results

4.3.1 Literature review

Eight cohort studies met the inclusion criteria (see Table 4.1). The reported rate ratios by age and sex from each study are presented in Table 4.6 in Appendix 4.A.

Study	Country	Period(s)	Reference
British Doctors	UK	1951–1961	(Doll and Hill, 1966)
		1951–1971	(Doll et al., 1994)
		1971–1991	(Doll et al., 1994)
		1951–1991	(Doll et al., 1994)
US Veterans	USA	1954–1962	(Kahn, 1966)
CPS-I	USA	1959–1972	(Burns et al., 1997c)
CPS-II	USA	1984–1988	(Thun and Heath, 1997)
Copenhagen Pop.	Denmark	1962–1994	(Prescott et al., 1998)
Framingham	USA	1972–1982	(Sorlie et al., 1989)
Kaiser Permanente	USA	1979–1988	(Friedman et al., 1997)
Three Communities	USA	1981–1988	(LaCroix et al., 1991)

Table 4.1: Study inclusions for a review of all cause mortality rate ratios by age and sex

There were several large cohort studies from Western populations that did not provide sufficient published details for inclusion in the review. In particular, results were not available from the 1966/68 NMFS/NHIS (Rosenbaum et al., 1998) and from the Nurses Health Study (Kawachi et al., 1997). Results were also incomplete from several of the studies included in the review. No results were available for former smokers from CPS-II. From the British Doctors Study, confidence intervals were not available for male former smokers and few results were available for female doctors.

A graphic comparison of the different rate ratios is given for male and female current smokers (Figures 4.1 and 4.2, respectively) and for male former smokers (Figure 4.3). A figure for female former smokers has not been shown as there were few data available. Pooled estimates from Holman et al. (1990) have been included, while estimates from English et al. (1995) were excluded due to broad age groups.

In general, there was considerable variability between and within studies. This can partially be explained by the inclusion of several smaller studies, including the Framingham Study and the Three Communities Study, which have fewer person-years of exposure and were less precise. Estimates for the younger age groups were quite imprecise in all studies.

The rate ratios for male current smokers tend to be higher in younger ages and decrease with age (Figure 4.1). Results from CPS-I suggest a decline in the two youngest age groups, however this is not apparent from CPS-II. There is a consistent upward shift from CPS-I to CPS-II, and from the earlier to the later set of results from the British Doctors Study.

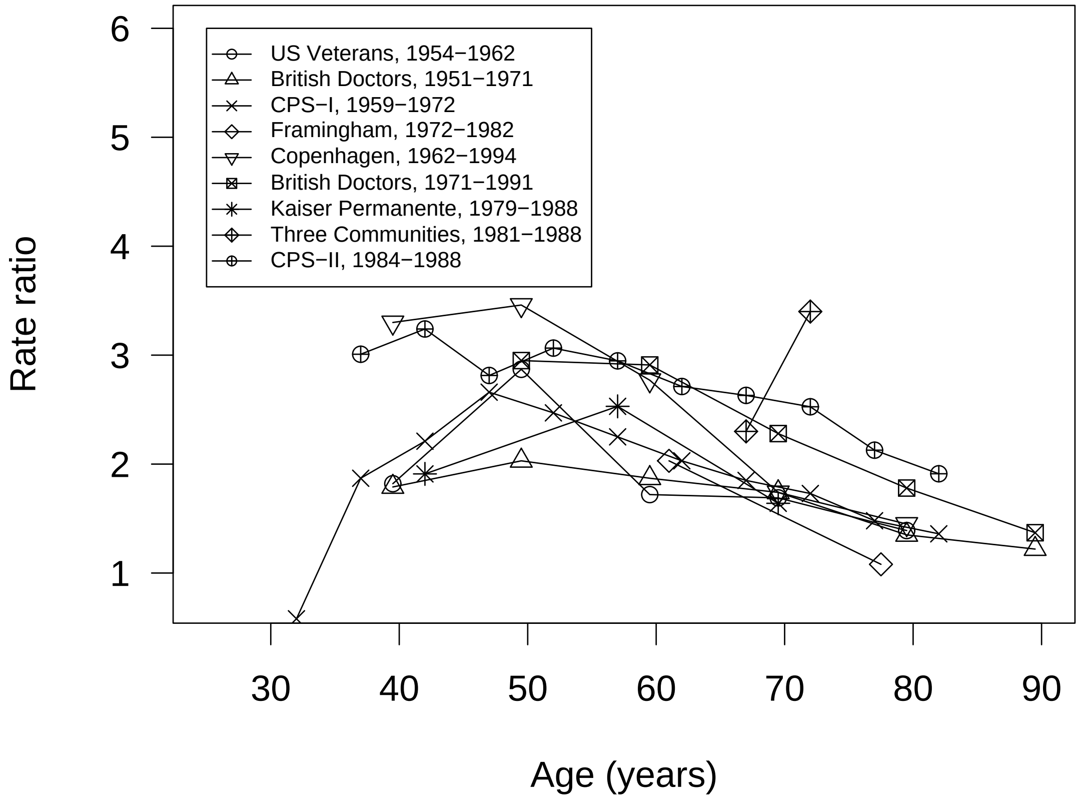


Figure 4.1: All cause mortality rate ratios for male current smokers

There were fewer studies reporting rate ratios for current smoking for females (Figure 4.2). The rate ratios were lower for females, which probably reflects lower exposure. There is a suggestion of an increase in age-specific rate ratios from CPS-I to CPS-II for older females. In contrast with male current smokers, the female current smokers exhibit curvature in their rate ratios, with lower rate ratios for younger and older women.

Aside from a slow downward trend by age, the rate ratios for male former smokers do not exhibit any strong patterns (Figure 4.3). The average rate ratio is around 1.4 for younger men and 1.2 for the older men.

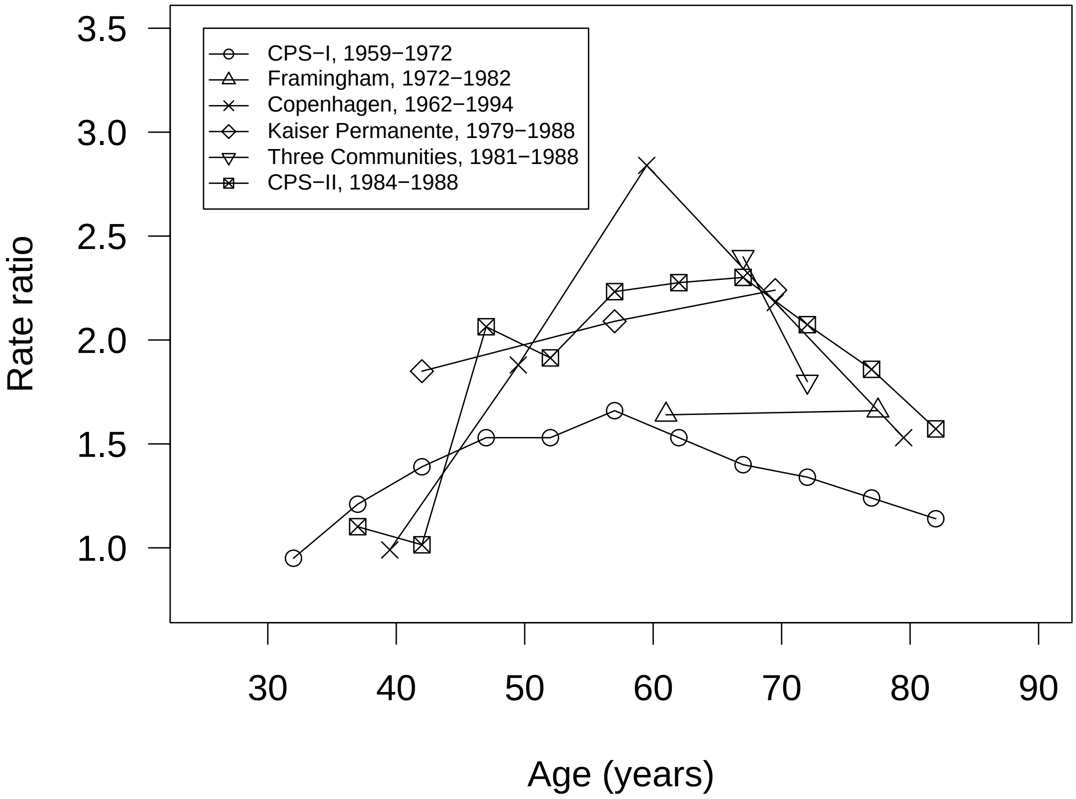


Figure 4.2: All cause mortality rate ratios for female current smoker

4.3.2 Meta-regression

Simple meta-regression which included age and study as a factor suggested that there were large differences between the studies after adjustment for age ($p < 0.0001$ for study heterogeneity for both males and females).

Estimates from the final models used in the formal meta-regression are shown in Tables 4.2 and 4.3. As expected, there was evidence of an increase over time in the rate ratios for current smokers in both males and females (see Figure 4.1 and 4.2). The proportional increase in rate ratios per year for females was greater than that for males. There was no corresponding evidence for a trend in rate ratios for former smokers.

Curvature of the log rate ratios was observed for female current smokers, while no such curvature was observed for male current smokers. Moreover, for current smokers, the rate ratios for males from CPS-I and CPS-II were systematically different from the other studies, while no such pattern was observed for females. This may be explained by there being fewer studies for females.

Male former smokers exhibited a gradual decline in rate ratios with age. For females, no other factors affected the rate ratios.

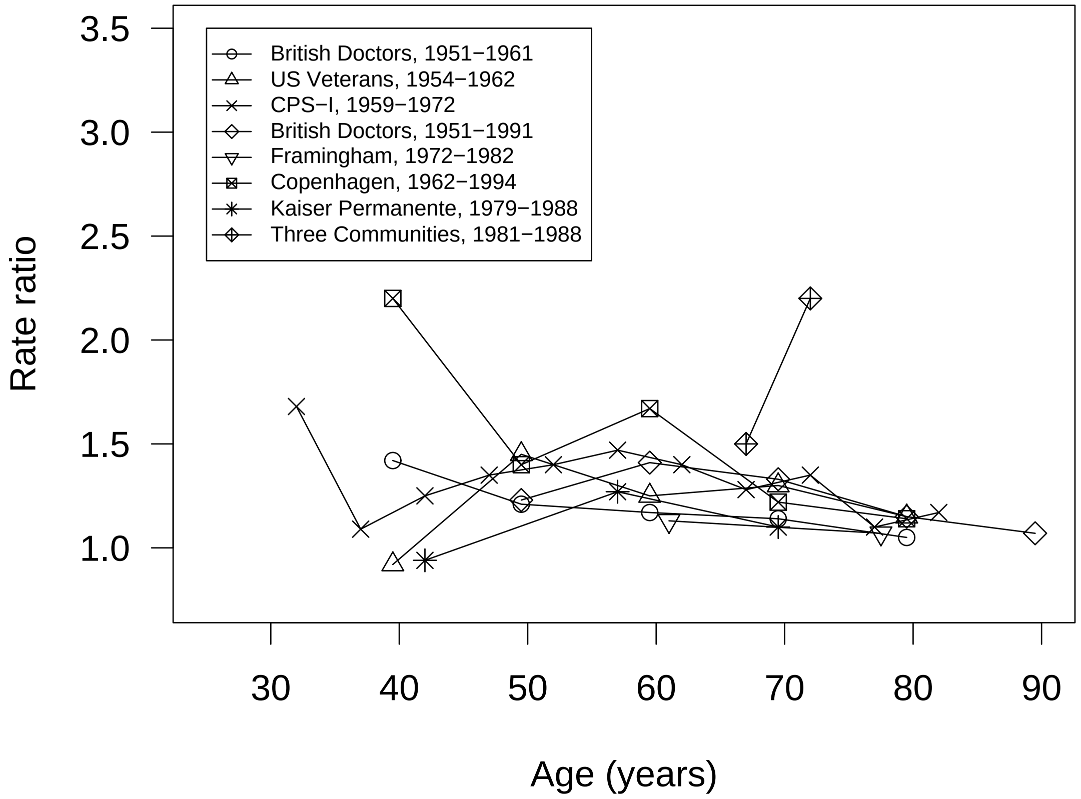


Figure 4.3: All cause mortality rate ratios for male former smokers

Smoking status	Sex	Rate ratio	95% CI
Current	Male	2.21	(1.93, 2.54)
	Female	2.29	(2.18, 2.41)
Former	Male	1.33	(1.28, 1.39)
	Female	1.11	(1.06, 1.16)

Table 4.2: Baseline all cause mortality rate ratios at age 60 in 1986, by sex and smoking status (meta-regression)

It is not proposed that these rate ratio equations be used for a particular population. Given the considerable heterogeneity, more population-specific information is required for valid estimation of rate ratios for a given population. This suggests that local estimates require an alternative approach to estimation. One such approach is following in the next section.

4.3.3 Estimates using cause-specific rate ratios

Using methods discussed in Section 4.2.4, prevalence proportions for 1966 were estimated using parameters given in Table 4.4 and Table 4.5. There was limited evidence

Smoking status	Sex	Variable	% change of rate ratio[a]	95% CI
Current	Male	Age	-1.65	(-1.97, -1.33)
		Calendar period	0.66	(0.12, 1.21)
		CPS-I	6.66	(-0.03, 13.79)
		CPS-II	29.40	(11.72, 49.87)
	Female	Age	-0.21	(-0.49, 0.09)
		Age^2	-0.064	(-0.085, -0.044)
		Calendar period	1.90	(1.63, 2.16)
Former	Male	Age	-0.65	(-0.99, -0.31)
	Female	(No other factors)	—	—

Table 4.3: Change in all cause mortality rate ratios relative to age 60 in 1986 for unit change in modelled factors, by sex and smoking status (meta-regression)

[a] The change in rate ratio was per unit change of a parameter. The unit of change for age, age^2 and calendar period was one year. Indicator variables were used to compare CPS-I and CPS-II with the other studies.

			log(OR)			
Country	Period	Sex	Constant	(Std Err)	Age	(Std Err)
Aust.	’74 to ’66	M	0	-	0.0084	(0.0014)
		F	0.025	(0.050)	0	-
NZ	’76 to ’66	M	0.123	(0.259)	0.0049	(0.0079)
		F	0.229	(0.241)	- 0.0097	(0.0073)
	’81 to ’86	M	0.027	(0.198)	- 0.0041	(0.0052)
		F	0.107	(0.164)	- 0.0036	(0.0045)

Table 4.4: Log odds ratios for the change in age-specific prevalence of current smokers, Australia and New Zealand

for a change in prevalence of current smoking, while there was good evidence for a change in the prevalence of former smokers over time. The point estimates for change were used for prediction, however the precision of such predictions was expected to be small.

All cause mortality rate ratios as estimated from cause-specific rate ratios and population-specific rates and prevalence for Australia are given in Figure 4.4. Results for Australia are tabulated in Table 4.9 in Appendix 4.A.

For the Australian population, there were marked differences over time, comparing estimates based on CPS-I with those based on CPS-II. This is consistent with the meta-regression. Moreover, there were marked differences between NMFS/NHIS and CPS-II for the period, particularly for males. This may be explained by the cohorts being representative for different populations which have different smoking

			log(OR)			
Country	Period	Sex	Constant	(Std Err)	Age	(Std Err)
Aust.	'74 to '66	M	-0.292	(0.077)	0	-
		F	-1.417	(0.463)	0.0247	(0.0131)
NZ	'76 to '66	M	0	-	0.0084	(0.0035)
		F	-1.625	(0.532)	0.0282	(0.0131)
	'81 to '86	M	0.107	(0.098)	0	-
		F	0	-	0.0052	(0.0020)

Table 4.5: Log odds ratios for the change in age-specific prevalence of former smokers, Australia and New Zealand

histories. The similarity at younger ages is a consequence of using the "other causes" from CPS-II in estimates for NMFS/NHIS.

For comparisons between populations, results for CPS-I and CPS-II are shown in Figure 4.5 and Figure 4.6, respectively. The estimates for Australia and New Zealand were remarkably similar, with New Zealand results being the same or higher by up to 3%. The CPS results were slightly higher on average than the Australian estimates (by 3–6%). The only major departures were between the country estimates and CPS at younger ages.

4.4 Discussion

There was considerable variation in all cause mortality rate ratios for former and current smoking by age, by sex, over time and between studies. The variation may be explained by differential exposure between populations, but this does not explain differences between studies within the same population, such as CPS-II and 1986 NMFS/NHIS.

There were several limitations to this analysis. First, the literature review cannot be considered exhaustive. However estimates from most of the large cohort studies over the previous 50 years were included, so that major changes in results by the inclusion of additional studies would be unlikely.

Second, there was a paucity of information available on females. This can be explained by the female smoking epidemic being later than that for men, so that early studies were either less likely to have included females or had fewer outcomes for females.

Third, rate ratios may be biased due to recall in retrospective estimates of smoking prevalence. Although evidence supports the validity of self-report of smoking for most study designs (see Chapter 1), some investigators warn that misclassification of

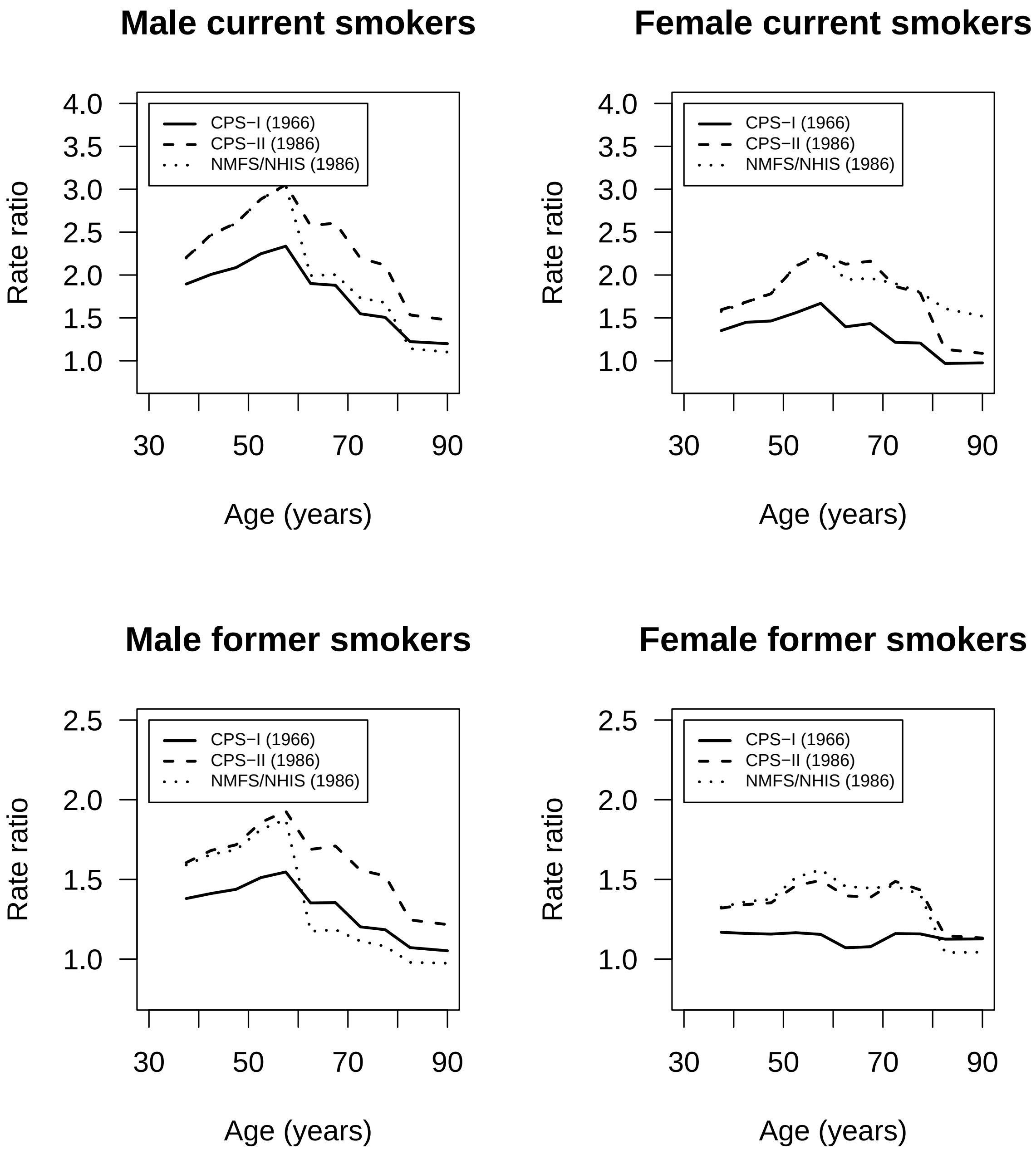


Figure 4.4: All cause mortality rate ratios for Australia estimated from cause-specific rate ratios from the given studies

former smokers as never smokers may bias rate ratios (van de Mheen and Gunning-Schepers, 1994, 1996).

Fourth, for the estimation that used cause-specific rate ratios, some components may have been inaccurate. In particular, prevalence estimates for 1966 were subject to random error due to small numbers and systematic error due to differential survival.

The meta-regression model assumed independent errors. An alternative approach

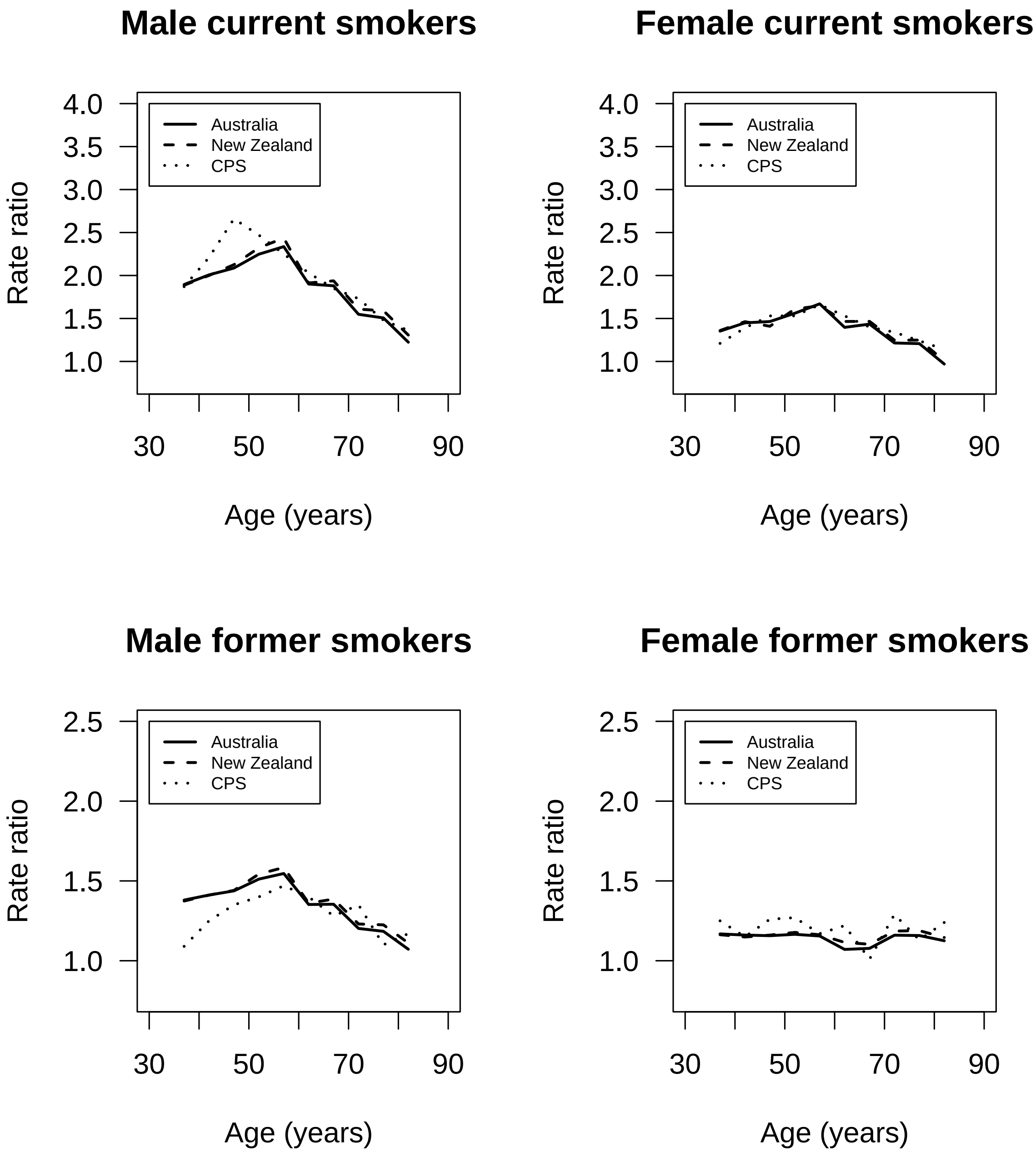


Figure 4.5: All cause mortality rate ratios for Australia and New Zealand for 1966 estimated from cause-specific rate ratios from CPS-I and population-specific descriptive data

would be to used a mixed model with errors within and between studies, which would lead to similar point estimates and wider confidence intervals. This would also allow for modelling of random effects between studies (Rothman and Greenland, 1998).

One consequence of the meta-regression is that a method is required to estimate valid all cause mortality rate ratios for New Zealand and Australia. The method using weighted cause-specific rate ratios which is insensitive to estimates of the cause-

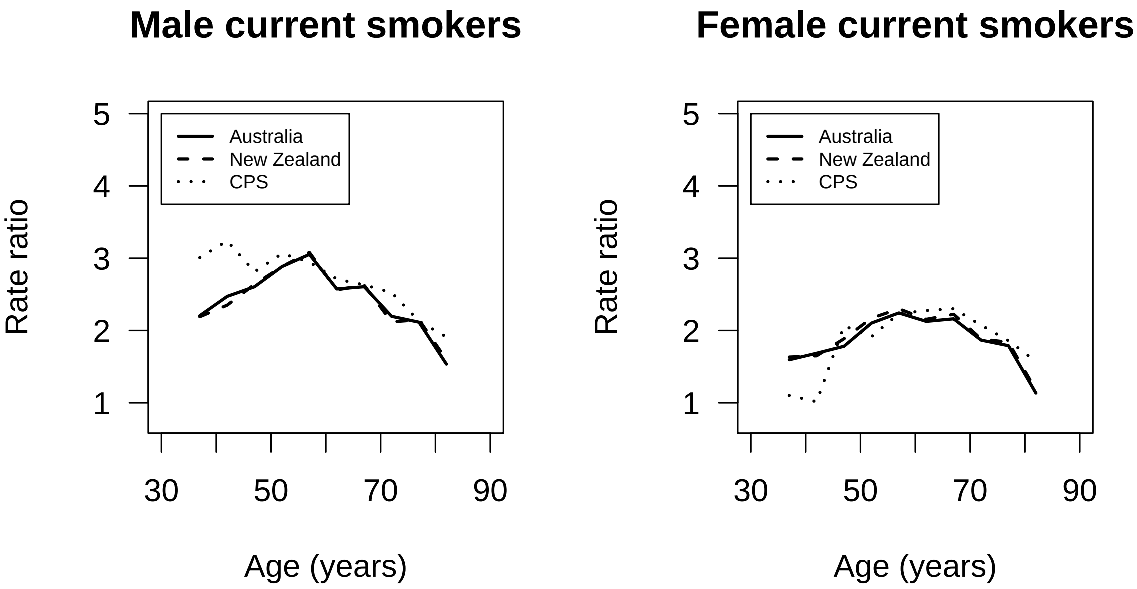


Figure 4.6: All cause mortality rate ratios for Australia and New Zealand for 1986 estimated from cause-specific rate ratios from CPS-II and population-specific descriptive data

specific rates, but is sensitive to large rate ratios (see Appendix 4.B on page 117). One unresolved concern is whether the cause-specific rate ratios from the large cohort studies are representative for New Zealand and Australia.

An alternative approach to estimate country-specific all cause mortality rate ratios would be to use population attributable fractions. However this approach would require either separate fractions for current and former smokers, or the specification of rate ratios for former smokers, hence it is not recommended.

All cause mortality rate ratio estimates for New Zealand and Australia are required for 1900–2050. As estimates using cause-specific rate ratios and local prevalence and rates were similar to observed all cause rate ratios, the use of observed rate ratios is considered parsimonious. Interpolation of age-specific rate ratios over time can be done on the log scale. Projections forward in time or predictions backward in time require greater care. Possible approaches include modelling for trend, modelling for a threshold, or keeping at the last known value. As there are only two studies being used (CPS-I and CPS-II), there is no means to model for trend or threshold. Following the last approach, values from CPS-I can be used for years prior to 1966 and values from CPS-II for years following 1986.

All cause mortality rate ratios are an expression of a variety of diseases, with different period, age and cohort effects affected by smoking exposure. There have been considerable changes in consumption and prevalence over time. Smoking exposure has been characterised as never, current or former smokers. This does not take account of dose, duration or pattern of smoking. Moreover, time since cessation is an important aetiological parameter. Given these aspects, it is not surprising that the effects of smoking are difficult to disentangle.

4.A Rate ratios

Table 4.6: All cause mortality rate ratios from the literature

			Current smokers		Ex smokers	
Study	Age	Sex	*RR*	95% CI	*RR*	95% CI
Holman et al	35-44	M	1.90	(1.59,2.27)	1.26	(0.98,1.63)
(Holman et al., 1990)	45-54	M	2.35	(2.16,2.55)	1.55	(1.39,1.73)
	55-64	M	1.78	(1.72,1.85)	1.33	(1.27,1.40)
	65-74	M	1.66	(1.61,1.72)	1.31	(1.26,1.37)
	75-84	M	1.33	(1.24,1.44)	1.18	(1.09,1.29)
English et al	<65	M	2.75	(2.53,2.98)	1.25	(1.15,1.35)
(English et al., 1995)	65+	M	1.16	(1.12,1.21)	1.10	(1.06,1.14)
	<65	F	1.75	(1.61,1.91)	1.24	(1.13,1.36)
	65+	F	1.65	(1.50,1.82)	1.03	(1.01,1.05)
British Doctors						
(1951–1991)	45-54	M	2.03	(1.69,2.43)	1.23	(–,–)
(Doll et al., 1994)	55-64	M	2.14	(1.90,2.41)	1.41	(–,–)
	65-74	M	1.98	(1.81,2.18)	1.33	(–,–)
	75-84	M	1.58	(1.44,1.73)	1.15	(–,–)
	85-94	M	1.30	(1.12,1.50)	1.07	(–,–)
(1951–1971)	35-44	M	1.79	(1.31,2.44)	–	(–,–)
(Doll et al., 1994)	45-54	M	2.03	(1.64,2.50)	–	(–,–)
	55-64	M	1.87	(1.58,2.21)	–	(–,–)
	65-74	M	1.74	(1.50,2.03)	–	(–,–)
	75-84	M	1.35	(1.16,1.56)	–	(–,–)
	85-94	M	1.22	(0.99,1.50)	–	(–,–)
(1971–1991)	45-54	M	2.95	(2.03,4.28)	–	(–,–)
(Doll et al., 1994)	55-64	M	2.91	(2.44,3.48)	–	(–,–)
	65-74	M	2.28	(2.01,2.57)	–	(–,–)
	75-84	M	1.78	(1.58,2.00)	–	(–,–)
	85-94	M	1.37	(1.11,1.68)	–	(–,–)
(1951–1961)	35-44	M	2.18	(1.37,3.48)	1.42	(0.71,2.84)
(Doll et al., 1966)	45-54	M	1.85	(1.34,2.55)	1.21	(0.79,1.84)
	55-64	M	1.71	(1.33,2.20)	1.17	(0.86,1.59)
	65-74	M	1.60	(1.26,2.03)	1.14	(0.86,1.51)

Table 4.6: (continued)

			Current smokers		Ex smokers	
Study	Age	Sex	*RR*	95% CI	*RR*	95% CI
	75-84	M	1.18	(0.96,1.45)	1.05	(0.83,1.33)
US Veterans	35-44	M	2.18	(1.37,3.48)	1.42	(0.71,2.84)
(Kahn, 1966)	45-54	M	1.85	(1.34,2.55)	1.21	(0.79,1.84)
	55-64	M	1.71	(1.33,2.20)	1.17	(0.86,1.59)
	65-74	M	1.60	(1.26,2.03)	1.14	(0.86,1.51)
	75-84	M	1.18	(0.96,1.45)	1.05	(0.83,1.33)
CPS-I	30-34	M	0.58	(0.20,1.67)	1.68	(0.34,8.33)
(Burns et al., 1997c)	35-39	M	1.87	(1.14,3.06)	1.09	(0.43,2.73)
	40-44	M	2.21	(1.69,2.88)	1.25	(0.79,1.97)
	45-49	M	2.66	(2.29,3.09)	1.35	(1.07,1.71)
	50-54	M	2.47	(2.27,2.69)	1.40	(1.23,1.58)
	55-59	M	2.25	(2.12,2.38)	1.47	(1.36,1.60)
	60-64	M	2.03	(1.94,2.13)	1.40	(1.31,1.49)
	65-69	M	1.85	(1.77,1.93)	1.28	(1.21,1.36)
	70-74	M	1.73	(1.66,1.81)	1.35	(1.28,1.43)
	75-79	M	1.48	(1.42,1.55)	1.10	(1.03,1.17)
	80-84	M	1.36	(1.28,1.45)	1.17	(1.08,1.27)
	85+	M	1.20	(1.11,1.31)	1.05	(0.95,1.17)
	30-34	F	0.95	(0.43,2.08)	3.02	(0.86,10.60)
	35-39	F	1.21	(0.88,1.67)	1.25	(0.60,2.59)
	40-44	F	1.39	(1.19,1.64)	1.15	(0.79,1.68)
	45-49	F	1.53	(1.39,1.68)	1.26	(1.02,1.55)
	50-54	F	1.53	(1.43,1.62)	1.27	(1.10,1.45)
	55-59	F	1.66	(1.57,1.74)	1.17	(1.03,1.32)
	60-64	F	1.53	(1.46,1.61)	1.22	(1.09,1.36)
	65-69	F	1.40	(1.34,1.47)	1.01	(0.90,1.14)
	70-74	F	1.34	(1.27,1.42)	1.29	(1.16,1.45)
	75-79	F	1.24	(1.17,1.32)	1.13	(1.00,1.28)
	80-84	F	1.14	(1.05,1.24)	1.24	(1.08,1.43)
	85+	F	0.92	(0.83,1.03)	1.12	(0.94,1.34)
CPS-II	35-39	M	3.01	(1.67,5.41)	–	(–,–)
(Thun and Heath, 1997)	40-44	M	3.24	(1.98,5.30)	–	(–,–)
	45-49	M	2.81	(2.24,3.53)	–	(–,–)

Table 4.6: (continued)

			Current smokers		Ex smokers	
Study	Age	Sex	*RR*	95% CI	*RR*	95% CI
	50-54	M	3.06	(2.68,3.51)	–	(–,–)
	55-59	M	2.95	(2.66,3.27)	–	(–,–)
	60-64	M	2.71	(2.50,2.94)	–	(–,–)
	65-69	M	2.63	(2.44,2.83)	–	(–,–)
	70-74	M	2.53	(2.35,2.71)	–	(–,–)
	75-79	M	2.13	(1.97,2.30)	–	(–,–)
	80-84	M	1.91	(1.71,2.14)	–	(–,–)
	85+	M	1.24	(1.03,1.48)	–	(–,–)
	35-39	F	1.10	(0.65,1.85)	–	(–,–)
	40-44	F	1.01	(0.72,1.43)	–	(–,–)
	45-49	F	2.06	(1.74,2.45)	–	(–,–)
	50-54	F	1.91	(1.70,2.16)	–	(–,–)
	55-59	F	2.23	(2.04,2.45)	–	(–,–)
	60-64	F	2.28	(2.10,2.46)	–	(–,–)
	65-69	F	2.30	(2.14,2.47)	–	(–,–)
	70-74	F	2.07	(1.93,2.23)	–	(–,–)
	75-79	F	1.86	(1.71,2.03)	–	(–,–)
	80-84	F	1.57	(1.39,1.78)	–	(–,–)
	85+	F	1.32	(1.13,1.54)	–	(–,–)
Copenhagen	35-44	M	3.30	(1.32,8.24)	2.20	(0.75,6.45)
(Prescott et al., 1998)	45-54	M	3.46	(2.19,5.46)	1.40	(0.83,2.37)
	55-64	M	2.77	(2.15,3.58)	1.67	(1.26,2.20)
	65-74	M	1.74	(1.47,2.06)	1.22	(1.02,1.46)
	75-84	M	1.45	(1.17,1.80)	1.14	(0.91,1.42)
	85+	M	1.27	(0.74,2.17)	0.97	(0.56,1.69)
	35-44	F	0.99	(0.49,2.03)	0.75	(0.25,2.23)
	45-54	F	1.88	(1.30,2.74)	1.07	(0.61,1.90)
	55-64	F	2.84	(2.19,3.70)	1.76	(1.25,2.47)
	65-74	F	2.18	(1.86,2.56)	1.35	(1.09,1.66)
	75-84	F	1.53	(1.31,1.77)	1.10	(0.91,1.33)
	85+	F	1.17	(0.89,1.52)	0.89	(0.64,1.24)
Framingham	53-69	M	2.03	(1.18,3.48)	1.13	(0.98,1.31)
(Sorlie et al., 1989)	70-85	M	1.08	(0.99,1.18)	1.07	(0.99,1.16)

Table 4.6: (continued)

			Current smokers		Ex smokers	
Study	Age	Sex	*RR*	95% CI	*RR*	95% CI
	53-69	F	1.64	(1.08,2.49)	1.27	(0.96,1.69)
	70-85	F	1.66	(1.13,2.44)	1.03	(0.99,1.07)
Kaiser Permanente	35-49	M	1.91	(1.31,2.76)	0.94	(0.56,1.59)
(Friedman et al., 1997)	50-64	M	2.53	(1.97,3.26)	1.27	(0.94,1.72)
	65-74	M	1.64	(1.28,2.11)	1.10	(0.87,1.41)
	75+	M	1.23	(0.91,1.66)	1.15	(0.92,1.44)
	35-49	F	1.85	(1.17,2.92)	0.58	(0.24,1.37)
	50-64	F	2.09	(1.61,2.73)	1.18	(0.82,1.70)
	65-74	F	2.24	(1.76,2.85)	1.18	(0.87,1.59)
	75+	F	1.36	(0.99,1.88)	1.36	(1.04,1.77)
Three Communities	65-69	M	2.30	(1.50,3.80)	1.50	(0.90,2.50)
(LaCroix et al., 1991)	70-74	M	3.40	(2.10,5.50)	2.20	(1.40,3.40)
	75+	M	1.30	(1.00,1.70)	1.10	(0.90,1.40)
	65-69	F	2.40	(1.50,3.80)	0.80	(0.40,1.70)
	70-74	F	1.80	(1.20,2.70)	1.60	(1.00,2.40)
	75+	F	1.20	(0.80,1.70)	0.90	(0.70,1.20)

Table 4.7: Cause-specific rate ratios, Cancer Prevention Study I

			Current smokers		Ex smokers	
Cause	Age	Sex	*RR*	95% CI	*RR*	95% CI
Lung cancer	35-59	M	12.76	(9.16,17.78)	2.79	(1.81,4.32)
	60-69	M	13.81	(10.87,17.54)	4.05	(3.04,5.39)
	70-79	M	13.77	(10.67,17.77)	4.63	(3.40,6.29)
	80+	M	5.68	(3.98,8.10)	3.06	(1.90,4.94)
	35-59	F	4.87	(3.87,6.12)	1.56	(0.87,2.79)
	60-69	F	4.13	(3.41,5.01)	1.21	(0.66,2.22)
	70-79	F	3.08	(2.35,4.04)	0.99	(0.41,2.41)
	80+	F	1.44	(0.70,2.96)	NA	
COPD	50-69	M	10.01	(7.29,13.74)	8.24	(5.83,11.64)
	70+	M	10.36	(8.20,13.09)	6.10	(4.66,7.98)

Table 4.7: (continued)

			Current smokers		Ex smokers	
Cause	Age	Sex	*RR*	95% CI	*RR*	95% CI
	50-69	F	5.81	(4.25,7.95)	3.93	(2.18,7.07)
	70+	F	4.33	(3.29,5.72)	2.93	(1.59,5.40)
CHD	35-59	M	2.53	(2.36,2.71)	1.67	(1.52,1.85)
	60-69	M	1.65	(1.57,1.73)	1.27	(1.19,1.36)
	70-79	M	1.32	(1.25,1.38)	1.13	(1.06,1.20)
	80+	M	1.12	(1.04,1.22)	1.09	(0.98,1.21)
	35-59	F	2.34	(2.13,2.57)	1.02	(0.79,1.31)
	60-69	F	1.53	(1.44,1.63)	0.94	(0.79,1.12)
	70-79	F	1.28	(1.20,1.37)	1.13	(0.98,1.31)
	80+	F	1.05	(0.95,1.17)	1.11	(0.91,1.34)
Cerebrovascular	35-59	M	2.18	(1.79,2.66)	1.22	(0.90,1.65)
	60-69	M	1.74	(1.55,1.96)	1.25	(1.06,1.47)
	70-79	M	1.15	(1.05,1.25)	0.94	(0.83,1.06)
	80+	M	0.96	(0.85,1.08)	0.80	(0.68,0.94)
	35-59	F	2.20	(1.91,2.53)	1.03	(0.71,1.50)
	60-69	F	1.59	(1.43,1.77)	1.31	(1.02,1.67)
	70-79	F	1.12	(1.01,1.24)	1.02	(0.82,1.28)
	80+	F	0.78	(0.67,0.91)	1.29	(1.03,1.61)
Other causes	35-59	M	1.74	(1.63,1.85)	1.34	(1.22,1.47)
	60-69	M	1.59	(1.52,1.67)	1.19	(1.11,1.28)
	70-79	M	1.46	(1.39,1.53)	1.17	(1.09,1.25)
	80+	M	1.18	(1.10,1.27)	1.04	(0.95,1.15)
	35-59	F	1.22	(1.17,1.27)	1.19	(1.09,1.31)
	60-69	F	1.19	(1.14,1.25)	1.06	(0.95,1.18)
	70-79	F	1.12	(1.06,1.19)	1.22	(1.09,1.37)
	80+	F	0.95	(0.86,1.05)	1.02	(0.85,1.21)

Table 4.8: Rate ratios for "other causes", Cancer Prevention Study II

		Current smokers		Ex smokers	
Age	Sex	*RR*	95% CI	*RR*	95% CI
35-59	M	2.05	(1.80,2.34)	1.58	(1.33,1.88)
60-69	M	2.00	(1.80,2.22)	1.50	(1.31,1.71)
70-79	M	1.83	(1.66,2.02)	1.47	(1.29,1.67)
80+	M	1.39	(1.18,1.65)	1.23	(1.00,1.51)
35-59	F	1.33	(1.20,1.47)	1.30	(1.12,1.50)
60-69	F	1.44	(1.32,1.57)	1.28	(1.11,1.48)
70-79	F	1.35	(1.22,1.48)	1.47	(1.25,1.73)
80+	F	1.24	(1.05,1.46)	1.33	(1.03,1.73)

Table 4.9: All cause mortality rate ratios for Australia

			Current smokers		Ex smokers	
Study	Age	Sex	*RR*	95% CI	*RR*	95% CI
CPS-I	35-39	M	1.89	(1.80,1.99)	1.38	(1.28,1.48)
(1966)	40-44	M	2.01	(1.91,2.10)	1.41	(1.31,1.51)
	45-49	M	2.09	(1.99,2.18)	1.44	(1.34,1.53)
	50-54	M	2.25	(2.14,2.36)	1.51	(1.41,1.61)
	55-59	M	2.34	(2.22,2.46)	1.55	(1.45,1.65)
	60-64	M	1.90	(1.82,1.98)	1.35	(1.29,1.42)
	65-69	M	1.88	(1.80,1.96)	1.35	(1.29,1.42)
	70-74	M	1.55	(1.49,1.61)	1.20	(1.15,1.25)
	75-79	M	1.51	(1.45,1.56)	1.18	(1.13,1.24)
	80-84	M	1.22	(1.16,1.28)	1.07	(1.00,1.14)
	85+	M	1.20	(1.14,1.26)	1.05	(0.99,1.12)
	35-39	F	1.35	(1.30,1.40)	1.17	(1.07,1.27)
	40-44	F	1.45	(1.39,1.50)	1.16	(1.06,1.26)
	45-49	F	1.46	(1.41,1.52)	1.16	(1.06,1.25)
	50-54	F	1.56	(1.50,1.62)	1.17	(1.07,1.26)
	55-59	F	1.67	(1.60,1.74)	1.15	(1.05,1.26)
	60-64	F	1.40	(1.35,1.45)	1.07	(0.98,1.16)
	65-69	F	1.43	(1.38,1.49)	1.08	(0.98,1.17)

Table 4.9: (continued)

			Current smokers		Ex smokers	
Study	Age	Sex	*RR*	95% CI	*RR*	95% CI
	70-74	F	1.21	(1.17,1.26)	1.16	(1.06,1.26)
	75-79	F	1.21	(1.16,1.26)	1.16	(1.06,1.25)
	80-84	F	0.97	(0.91,1.03)	1.13	(1.00,1.25)
	85+	F	0.98	(0.91,1.04)	1.13	(1.00,1.25)
CPS-II	35-39	M	2.20	(1.97,2.44)	1.61	(1.37,1.84)
(1986)	40-44	M	2.47	(2.23,2.72)	1.68	(1.46,1.90)
	45-49	M	2.61	(2.35,2.86)	1.72	(1.51,1.92)
	50-54	M	2.88	(2.58,3.18)	1.86	(1.64,2.08)
	55-59	M	3.05	(2.72,3.39)	1.93	(1.70,2.15)
	60-64	M	2.57	(2.35,2.80)	1.69	(1.55,1.83)
	65-69	M	2.61	(2.38,2.83)	1.71	(1.56,1.86)
	70-74	M	2.20	(2.02,2.37)	1.56	(1.44,1.67)
	75-79	M	2.11	(1.96,2.27)	1.52	(1.41,1.63)
	80-84	M	1.53	(1.37,1.70)	1.25	(1.11,1.38)
	85+	M	1.48	(1.32,1.64)	1.22	(1.08,1.36)
	35-39	F	1.60	(1.45,1.74)	1.32	(1.15,1.49)
	40-44	F	1.69	(1.54,1.83)	1.34	(1.17,1.52)
	45-49	F	1.78	(1.63,1.94)	1.35	(1.19,1.52)
	50-54	F	2.10	(1.91,2.30)	1.46	(1.29,1.63)
	55-59	F	2.24	(2.03,2.46)	1.49	(1.32,1.67)
	60-64	F	2.13	(1.98,2.27)	1.40	(1.26,1.54)
	65-69	F	2.16	(2.01,2.31)	1.39	(1.25,1.52)
	70-74	F	1.87	(1.74,1.99)	1.49	(1.35,1.63)
	75-79	F	1.79	(1.67,1.91)	1.43	(1.30,1.57)
	80-84	F	1.13	(1.01,1.26)	1.15	(0.98,1.31)
	85+	F	1.09	(0.96,1.21)	1.13	(0.96,1.31)
NMFS/NHIS	35-39	M	2.20	(1.92,2.48)	1.59	(1.34,1.84)
(1986)	40-44	M	2.46	(2.04,2.88)	1.66	(1.38,1.93)
	45-49	M	2.60	(2.08,3.12)	1.69	(1.38,1.99)
	50-54	M	2.89	(2.18,3.59)	1.81	(1.45,2.18)
	55-59	M	3.04	(2.21,3.88)	1.88	(1.46,2.30)
	60-64	M	1.99	(1.56,2.43)	1.17	(0.95,1.39)
	65-69	M	2.00	(1.55,2.46)	1.18	(0.94,1.43)

Table 4.9: (continued)

			Current smokers		Ex smokers	
Study	Age	Sex	*RR*	95% CI	*RR*	95% CI
	70-74	M	1.73	(1.42,2.05)	1.11	(0.95,1.28)
	75-79	M	1.68	(1.39,1.97)	1.08	(0.92,1.24)
	80-84	M	1.14	(0.88,1.40)	0.98	(0.81,1.15)
	85+	M	1.10	(0.89,1.32)	0.97	(0.81,1.13)
	35-39	F	1.58	(1.36,1.79)	1.33	(1.13,1.52)
	40-44	F	1.68	(1.38,1.98)	1.36	(1.15,1.57)
	45-49	F	1.79	(1.45,2.13)	1.38	(1.15,1.60)
	50-54	F	2.10	(1.62,2.58)	1.51	(1.20,1.83)
	55-59	F	2.27	(1.68,2.86)	1.56	(1.20,1.92)
	60-64	F	1.94	(1.56,2.33)	1.46	(1.15,1.77)
	65-69	F	1.96	(1.55,2.37)	1.45	(1.12,1.77)
	70-74	F	1.90	(1.58,2.22)	1.46	(1.24,1.67)
	75-79	F	1.80	(1.50,2.10)	1.41	(1.20,1.62)
	80-84	F	1.61	(1.26,1.96)	1.04	(0.84,1.24)
	85+	F	1.52	(1.19,1.85)	1.04	(0.84,1.25)

4.B Sensitivity analysis for estimates based on cause-specific rate ratios

A simple decomposition of RR_c and RR_x between the different causes is offered by Equations (4.4) and (4.5), giving the ability to ascertain which causes are explaining changes in the all cause mortality rate ratios.

The estimates may be sensitive to values of the different parameters, either due to variation in estimation of parameters or parameters changing over time. For changes in cause-specific mortality rates for never smokers, we have the following differential equation for RR (being either RR_c or RR_x):

$$\frac{\mathrm{d}RR}{\mathrm{d}\mu_{0,i}} = \frac{(1-w_i)RR_i}{\mu_0}$$

so that $\delta RR = \delta \times w_i(1-w_i)\, RR_i$ which is a small change in the rate ratio for small w_i or small RR_i. As the majority of the death rate among never smokers is for diseases that are not strongly associated with smoking, then the rate ratio is relatively insensitive.

For changes in a cause-specific rate ratio, we find

$$\frac{\mathrm{d}RR}{\mathrm{d}RR_i} = w_i RR_i$$

so that $\delta RR = \delta \times w_i RR_i^2$. This suggests that where the cause-specific rate ratios are large, such as for lung cancer, care may be required. However, the rate ratios may not be sensitive to the rate ratios for lung cancer as lung cancer is a rare disease among never smokers, explaining a small proportion of never smoker mortality.

The analysis for the change in RR with respect to a change in smoking prevalence is more involved. As there are three smoking states, any change in one state can affect either or both of the other two states. For under-reporting of smoking prevalence, the prevalence of current and former smokers would both decline, and in the following it has been assumed that the ratio of the prevalence for the two groups would remain constant. It is convenient to define the population attributable fraction for cause i as $PAF_i = (\mu_i - \mu_{0,i})/\mu_i$. Then

$$\frac{\mathrm{d}RR}{\mathrm{d}\pi_c} = \frac{1}{\pi_c} \sum_{i,j} RR_i w_i w_j (PAF_j - PAF_i)$$

so that $\delta RR = \delta \times \sum_{i,j} RR_i w_i w_j (PAF_j - PAF_i)$. The sum, when restricted to causes with similar RR_i or causes with small population attributable fractions, will be close to zero. Consideration of the other combinations suggests the approximation for the change in RR given a relative change δ in prevalence:

$$\delta RR \approx -\delta \times \sum_i (RR_i - 1) w_i (1 - w_i) PAF_i,$$

which is likely to be small. A more simple interpretation of this is that a change in prevalence will alter the weights w_i, which will not affect appreciably the weighted estimate.

Chapter 5

Smoking initiation and cessation rates using retrospective data

Abstract

Aim: To use recall from survey data to estimate smoking initiation and cessation rates. **Methods:** Data were taken from the Risk Factor Prevalence Study, the 1998 National Drugs Strategy Household Survey and the 1996/97 New Zealand National Health Survey. Modern regression methods were used to smooth the steps in the cumulative hazard function to estimate the hazard. For smoothing over the Lexis diagram, a transformation was used to adjust for triangular boundaries, and smoothing was performed for both the risk set and the events. Adjustment was made for bias due to differential survival. **Results:** Differences in the patterns of change in smoking prevalence by sex were explained by differential uptake and cessation over the Lexis diagram. Older cohorts exhibited uptake across a wider range of ages. **Discussion:** A useful method for estimation of rates of change for early periods has been described. There are concerns with the validity of the retrospective data on which this analysis is based.

5.1 Introduction

Two general approaches have been used recently to estimate smoking cessation rates. The first approach is to use current status data. A number of countries, including Australia and New Zealand, have limited reliable cross-sectional prevalence data before the 1970s, which is an important data constraint. For applications of this approach using recent survey data, see Chapters 6 and 7. The second approach is to estimate the hazard using retrospective data and survival analysis methods, which is the focus of the current chapter.

5.1.1 Definitions

To review definitions from Chapter 3, *retrospective* data are defined as data collected from a cross-sectional survey based on recall of smoking uptake and cessation. It is possible to obtain estimates from multiple cross-sectional surveys and make comparisons between them, however only the case of a single survey or pooled surveys will be considered here. *Current status* data are defined as data based on the current state at time of interview.

The dynamics of interest are the rates of change in the cumulative measures. In survival analysis, these rates are usually termed the *hazard rates*, while for multistate models they are often called *transition intensities*. Important qualities of retrospective data are the existence of *right censoring*, where an individual may not have started (or stopped) smoking by the time of interview, and *left truncation*, where an individual cannot begin to stop smoking until they have started smoking.

5.1.2 Previous research

There is an established literature describing smoking behaviour using retrospective data. Retrospective synthetic cohort analysis for smoking was popularised by Harris (1980, 1983). Amongst other methodological contributions, Christie et al. (1986) gave an early presentation of cumulative measures of ever and former smoking and Birkett (1997) used proportional hazards models to compare smoking uptake between birth cohorts.

The state of art for retrospective smoking analysis is currently represented by Burns et al. (1997b) who used over 450,000 observations from the National Health Interview Surveys for 1965–1991. US estimates for the prevalence of ever and current smokers were adjusted for differential mortality and were stratified by birth cohort. Age was the main time scale, with data presented either by cohort or by period. There were marked differences in smoking behaviour by sex and across time. Males born late in the nineteenth century had moderate levels of smoking, with ever smoking prevalence peaking for those born during 1910–1914 and then declining with increasing rapidity through the twentieth century. In contrast, females born in 1885–1889 had very low levels of smoking, with ever smoking peaking for those born in 1940–1944, and then an irregular decline since that time. Actuarial estimates of cessation rates were also presented, albeit as a set of irregular lines. The cessation hazards could have been estimated using modern survival methods, however it is necessary first to address some technical difficulties.

5.1.3 Technical difficulties

There are four technical difficulties when estimating smoking initiation and cessation rates using retrospective data. First, the hazard rates may be *heterogeneous* over time. In particular, there is anecdotal evidence that the proportionality of the hazards by birth cohorts may not hold (Burns et al., 1997b). Given the limited investigation about the hazards, appropriate parametric forms are not known. The main implication is the need for non-parametric hazard estimation over the Lexis diagram. Keiding (1990) has reviewed the statistical aspects of hazard estimation over the Lexis diagram, however the methods have received relatively little application over the past decade. Heterogeneity of the hazards is also an issue for estimation using current status data.

Second, *differential survival* is known to bias retrospective estimates of smoking prevalence (see Section 3.6.2.7 and Harris, 1983). Hazard estimates are also biased by differential mortality (see Section 3.6.2.8 and Keiding, 1990, 1991; Elandt-Johnson and Johnson, 1980). Observed initiation rate estimates will tend to be lower than the true level, because fewer ever smokers survive relative to never smokers, and observed cessation rates will tend to be higher than the true level, because more former than current smokers will survive.

Third, the *validity* of retrospective estimates compared with current status data is generally open to question (Diamond and McDonald, 1991). An article by van de Mheen and (1994), suggesting that some former smokers change their recall of being ever smokers over time, has received little attention in studies which estimate retrospective smoking prevalence. Assessment of the validity is postponed until Chapter 6.

Fourth, there is a dearth of *software* available for hazard estimation of left truncated and right censored weighted data on the Lexis diagram.

5.1.4 Time scales

Age was used as the main time scale for initiation given their close relationship. However the choice of time scale for cessation is more equivocal, where either age or duration could be used. For more recent birth cohorts, age and duration are closely correlated because initiation has tended to take place over a restricted age range. Duration would have been advantageous for estimation, as the data would not have been left truncated and standard software could then have been used.

However the use of age was more flexible, allowing for Markov models and, by the inclusion of either cohort or period as a covariate, allowing for estimation over the Lexis diagram. For a related discussion of the main time scale, see Korn et al. (1997). Andersen et al. (1993, Chapter 10) provide a technical overview for analysis on multiple time scales.

An analysis by cohort and age allows the event history for an individual to be compactly represented. In contrast, an analysis by age and period requires a decomposition of the person-years by age and period. Although care is required in this counting process, demographers and epidemiologists have commonly used this procedure for cohort analysis. Software tools are readily available, including macros and functions in `SAS`, `Stata` and `R/S-Plus`.

Any joint modelling of age, period and cohort leads to the familiar problem of non-identifiability, where the three time scales are linearly dependent. This is not a problem for the current research question, as a non-parametric curve has been fitted across the Lexis diagram so that age-cohort and age-period models give similar results: the models are two ways of expressing the same surface. This expresses the geometric difficulty of parametric age-period-cohort modelling where the surface cannot be easily decomposed into effects due to the three time scales. Some authors simply suggest that age-period-cohort modelling should not be attempted (Moolgavkar et al., 1998).

One approach to the choice of time scales is to consider the likely influences due to age, period and cohort. Period is important where there are important changes in societal attitudes to smoking behaviour. Smoking initiation increased across ages after World War II and there have been recent declines in prevalence across all ages, suggesting period effects. Age has a strong relationship with uptake. Moreover, at older ages smokers may experience increasing morbidity and mortality by their peers or get sick themselves, leading to increased quitting. Cohort effects may also be seen in mortality outcomes, where a particular birth cohort may be younger during a period of more rapid smoking uptake. However, as uptake has most commonly been restricted to younger ages, period and age may explain any cohort effect. Moreover, rates of behavioural change tend to express recent (period) changes in society. However, in reality, there is a complicated interaction between the three time scales.

5.1.5 Outline

For this chapter, smoking initiation and cessation rates are estimated using data from the Risk Factor Prevalence Study held during the 1980s and from two additional Australasian surveys. Cumulative rates and hazard rates are presented stratified by birth cohort. The hazard rates are smoothed using kernel methods. Rates are estimated over the Lexis diagram (see Figure 5.1) using novel methods employing local likelihood density estimation together with a suitable age-cohort transformation. Finally, some investigation is made using semi-parametric hazard models to suggest hazards for earlier cohorts.

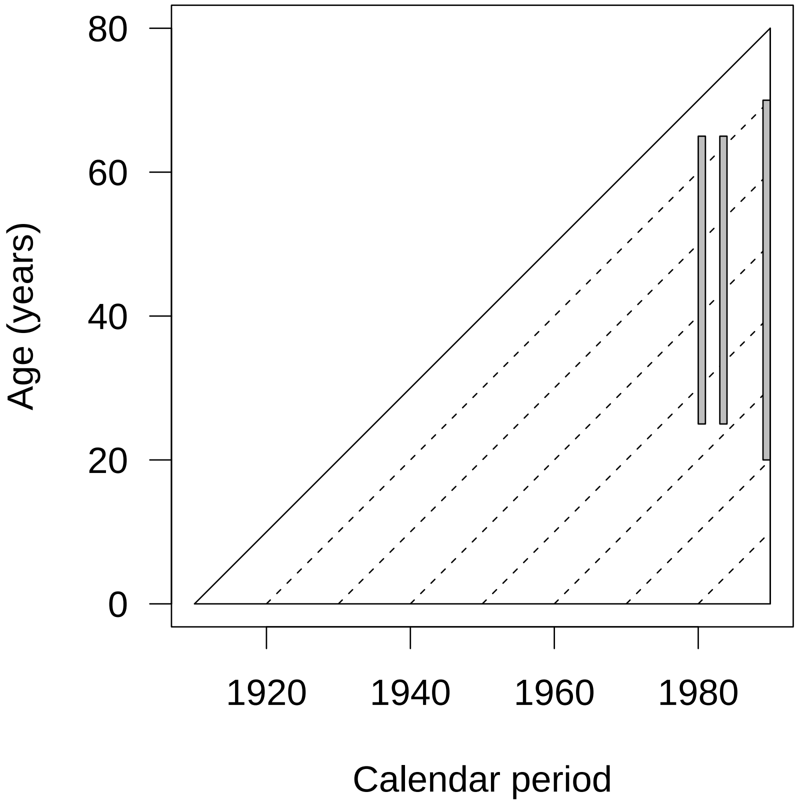


Figure 5.1: Lexis diagram for retrospective recall of smoking from the Risk Factor Prevalence Study (birth cohorts divided by dashed lines; surveys shown in grey)

5.2 Methods

5.2.1 Data

Retrospective smoking initiation and cessation were estimated from the Risk Factor Prevalence Study held during 1980, 1983 and 1989. In addition, estimates have been presented from the 1998 National Drugs Strategy Household Survey for Australia and from the 1996/97 New Zealand Health Survey.

To avoid inclusion of short-term quitting or smoking uptake, the three years prior to interview were not included in any analysis of smoking initiation or uptake. Smoking cessation is thus defined as having stopped smoking for at least three years. Due to numerical instability, the rare examples of cessation before 14 years of age were considered experimentation and were ignored.

After pre-processing, there was a record for each respondent including variables for: effective cell weight; cohort and sex; age of smoking initiation together with an associated event variable; and age of smoking cessation together with an associated event variable.

For details about the data sources and their general data processing, see Chapter 2.

5.2.2 Statistical methods

The following analyses were performed separately for each sex. For analytical details, see Chapter 3.

5.2.2.1 Cumulative hazards

Survival was estimated using the Kaplan-Meier estimator for data stratified by ten-year birth cohorts (Klein and Moeschberger, 1997). The estimation used the `survfit()` function from the `survival5` library (Therneau and Grambsch, 2000) in `R` software. Variance was estimated using Greenwood's formula. The point estimate for the cumulative hazard was estimated from the log of survival. The variance was calculated by the width of the 95% confidence interval divided by 2×1.96. Steps were calculated for the change in the point estimate and the variance of the cumulative hazard. The *slope* of the cumulative hazard represents a measure of the underlying hazard.

5.2.2.2 Hazard for a given birth cohort

The hazard was estimated by smoothing the steps in the cumulative hazard. Smoothing was performed using an Epanechnikov kernel with modifications at the boundaries as proposed by Müller and Wang (1994). Variance was estimated by smoothing the change in variance of the cumulative hazard and the bandwidth was estimated using generalised cross validation (Klein and Moeschberger, 1997). The bandwidth for initiation and cessation were taken as being 7 and 8 years, respectively. The variance is for the *smoothed* estimator rather than the process itself. Estimation was performed by single year birth cohorts using `R` software.

The results are presented using conditional plots, with separate plots conditional upon a range of years for the birth cohorts (Cleveland, 1993). The numbers of events were used to calculate overlapping sets of cohorts, as described by the upper conditioning panel. The earliest cohort is presented at the bottom left of the result plots, proceeding to later cohorts by moving from left to right across the rows and then up to the next row, finishing at the top right. Instability due to small numbers may influence the cessation hazard estimation, so ages with fewer than ten respondents at risk were excluded. It should be noted that the cohort intervals were different from those used for the cumulative hazard analysis because the hazard estimates were less precise.

5.2.2.3 Hazards across the Lexis diagram

The method used to smooth the hazard across the Lexis diagram is described by Keiding (1990). The estimation involved smoothing first the risk set and second the ratio of steps in the counting process and the smoothed risk set. Each smoothing was performed in two dimensions.

Hazard estimation at boundaries was dealt with using a novel transformation. For a set of ages and cohorts restricted to a triangle, the set was transformed to a square. For the square, the risk set and the steps of the ratio on the square were estimated. Then the set was back-transformed.

The smoothing was performed using weighted local likelihood density estimation, scaled by the total cumulative hazard to estimate the hazard function. The `locfit` package for `R` statistical software was used. The smoothing parameter in the `locfit()` function was estimated using generalised cross-validation as described by Loader (1999).

The algorithm was validated using a simulation data set (see Appendix 5.A).

Details of these estimation methods are given in Chapter 3.

5.2.2.4 Proportional hazards models

A proportional hazards model was used with cohort as the only covariate, with the functional form determined using penalised splines and polynomial models (Therneau and Gram 2000). A global test for proportionality was carried out using the `cox.zph()` function from the `survival5` library in `R`. The test was carried out for each sex, both for all cohorts and for those cohorts born before 1950. The birth cohorts were divided at 1950 based upon inspection of the data and observed differences for initiation and cessation before and after this year.

The change in likelihood was used for comparisons between the penalised spline models and parametric models. The comparisons were less formal for the spline models as some models were not strictly nested.

5.2.2.5 Adjustment for differential mortality

Results were adjusted for differential mortality using survival ratios estimated in Chapter 8. To outline the estimation of the survival ratios, unadjusted retrospective hazard estimates were combined with current status hazard estimates, mortality rates and rate ratios, and initial prevalence. This fully specified a multi-state smoking model that was pushed back in time, iteratively estimating state parameters and adjusting the retrospective hazard estimates for differential mortality. Survival estimates and their ratios were calculated from products of the transition probabilities matrices.

The survival ratios for never to current smokers and for current to former smokers need to be defined for a range of times from age of event to age of interview. These ratios were predicted using local likelihood using a log link.

The survival ratios for never to current smokers were used as multipliers to adjust the smoking initiation rates. Similarly, smoking cessation rates were adjusted by multiplying by the survival ratios for current to former smokers. The adjusted hazard estimates were used to calculate the adjusted cumulative hazards.

The proportional hazards models require sample weights to be adjusted for differential selection. For smoking initiation, the individuals who began smoking were weighted by the survival ratio of never to current smokers moving from their time of initiation to their age at interview. Although group changes for cessation were accounted for in the survival ratio estimation, no account was taken for whether an individual later quit smoking. For smoking cessation, those that began smoking had sample weights adjusted for the survival ratio between current and former smokers.

For technical detail, see Chapter 3.

5.3 Results

5.3.1 Smoking initiation from the Risk Factor Prevalence Study

5.3.1.1 Cumulative hazard

Cumulative hazards[1] for smoking initiation are shown in Figure 5.2. Male cumulative hazard estimates from the Risk Factor Prevalence Study peaked at about 1.3 for the 1920–1929 birth cohort, which was a prevalence of ever smoking of approximately 73% $(= 1 - \exp(-1.3))$. For more recent birth cohorts, there was a consistent decline in the cumulative hazard of ever smoking, down to approximately 45% of the population at risk for the most recent birth cohort (1960–1969).

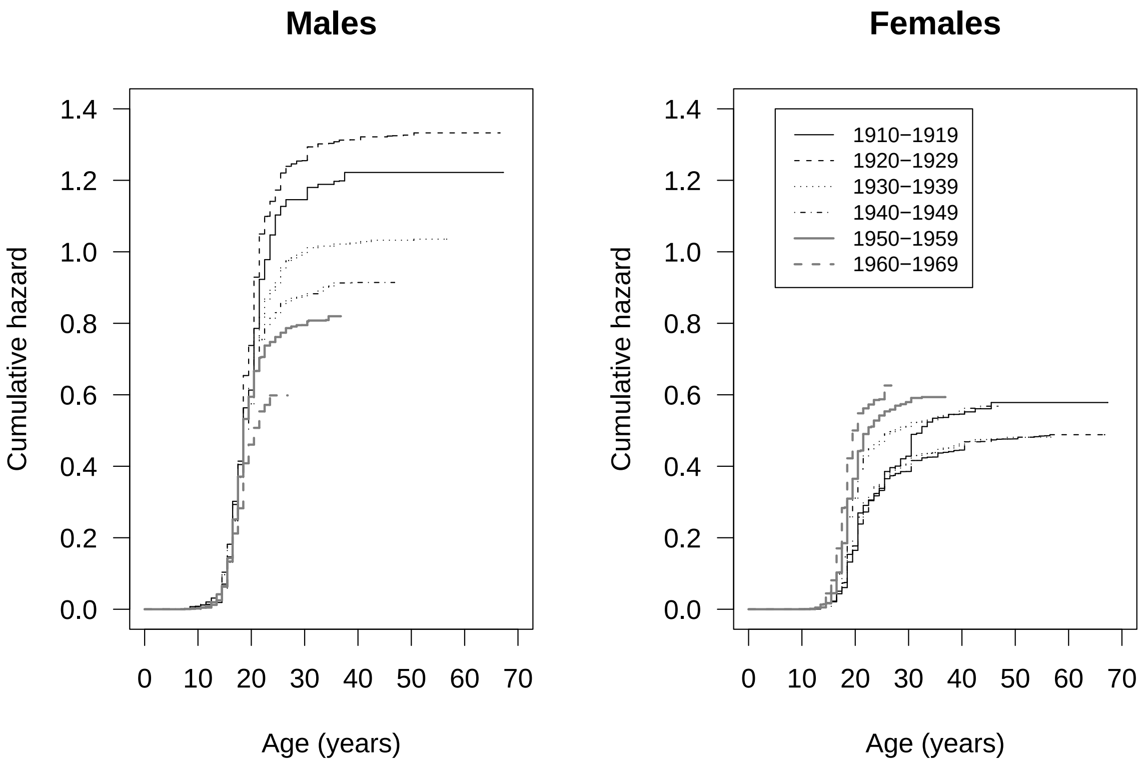


Figure 5.2: Cumulative hazard of initiation of tobacco smoking from the Risk Factor Prevalence Study, by sex

The pattern for females was substantially different. The oldest cohorts have a

[1] Cumulative hazards have been reported here because the slope is interpretable as the hazard for both initiation and cessation. In contrast, "survival" for smoking cessation does not have a ready interpretation because smoking is taken up at different ages. A presentation of the cumulative hazards using colour would improve interpretation.

prevalence of ever smoking of 35–40%, while more recent births cohorts experienced increasing smoking uptake, with the most recent birth cohort being at 45% or higher. The pattern of uptake for the younger cohorts is similar to that for males of the same cohort.

Uptake also varied between the two sexes and by birth cohort. Males of all cohorts tended to experience a rapid uptake in the late teens and early to mid twenties. The pattern was relatively consistent between the birth cohorts, providing qualitative evidence for proportionality of the underlying hazard. In contrast, the older female cohorts experienced negligible uptake before 20 years of age, with gradual uptake though the twenties and thirties. The two youngest female birth cohorts followed a pattern that was similar to males of comparable year of birth.

5.3.1.2 Hazard estimates

For the conditional plots of smoking initiation for males, more recent birth cohorts had lower rates than the older birth cohorts (see Figure 5.4 on page 130; see also page 124 for an explanation of conditional plots). The general pattern was for rates to be low before 10 years of age, rising during the early teenage years to peak at 18–20 years of age. After that age, the rates declined rapidly by the mid twenties and tended to be low after 30 years of age. The rates after age 30 years are less than 2% since 1950. Peak initiation shifted from around 20 years of age in 1940 to around 18 years of age since 1950.

The rise and fall of smoking cessation for females was similar to that for males (Figure 5.3 on page 129). The younger birth cohorts had the highest levels of smoking uptake, where the oldest birth cohort had the lowest (unadjusted) level of smoking initiation. Similar to the males, the older female birth cohorts exhibited a spread of initiation into older ages.

Initiation rates can also be represented as a surface on the Lexis diagram (see Figure 5.5). The contour plots emphasise the spread in uptake for the earlier cohorts, particularly for females where the contour stays at 0.02 per year for ages 30–35 years during 1950. The spread may be due to a period effect during and after World War II, which is also recognised from the general increase of tobacco consumption at that time.

5.3.1.3 Proportional hazards models

An inspection of the conditional plots suggests some differences in the shape of the age-specific hazard for initiation. For proportional hazards models that used

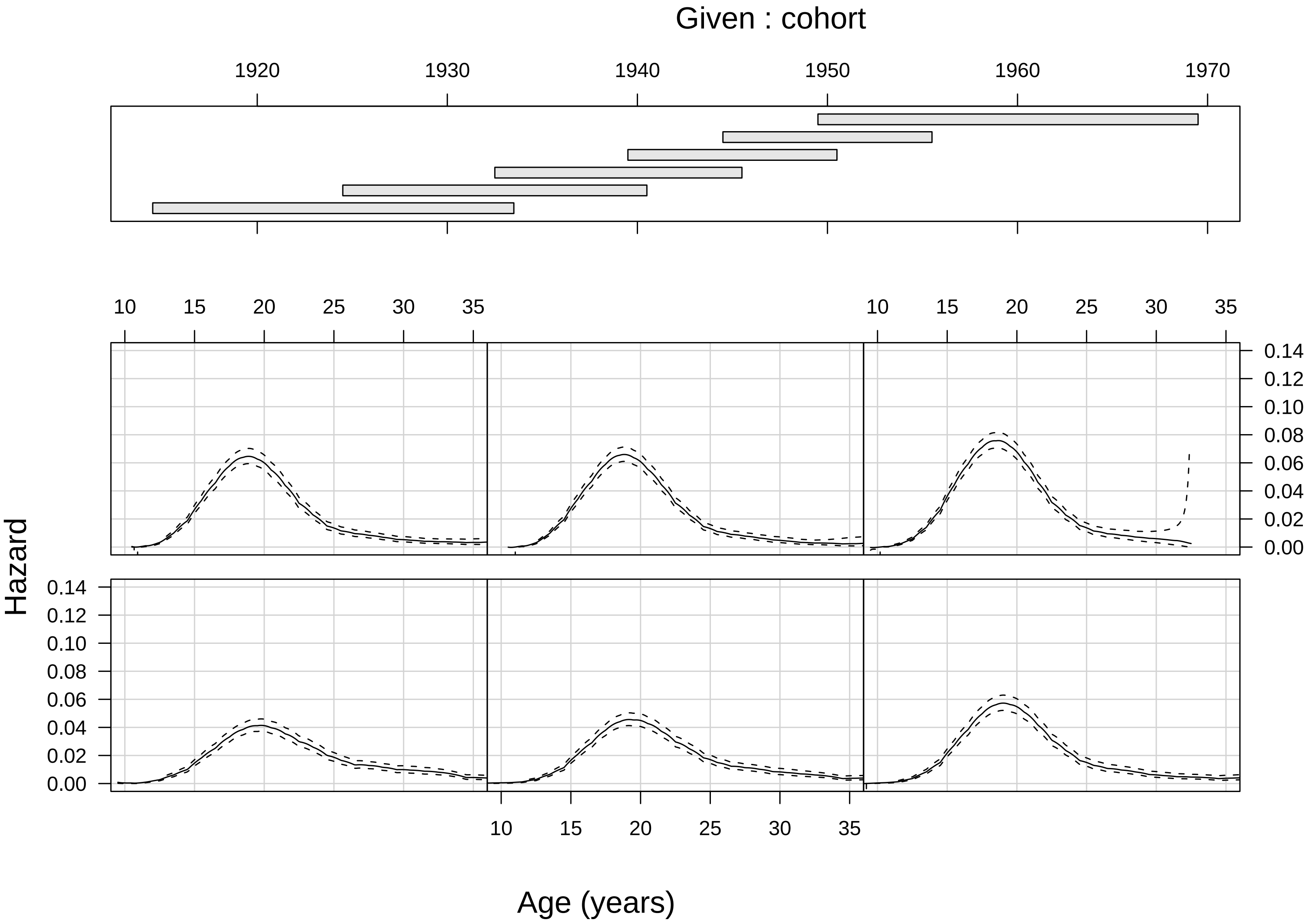


Figure 5.3: Coplot of initiation of tobacco smoking for females from the Risk Factor Prevalence Study

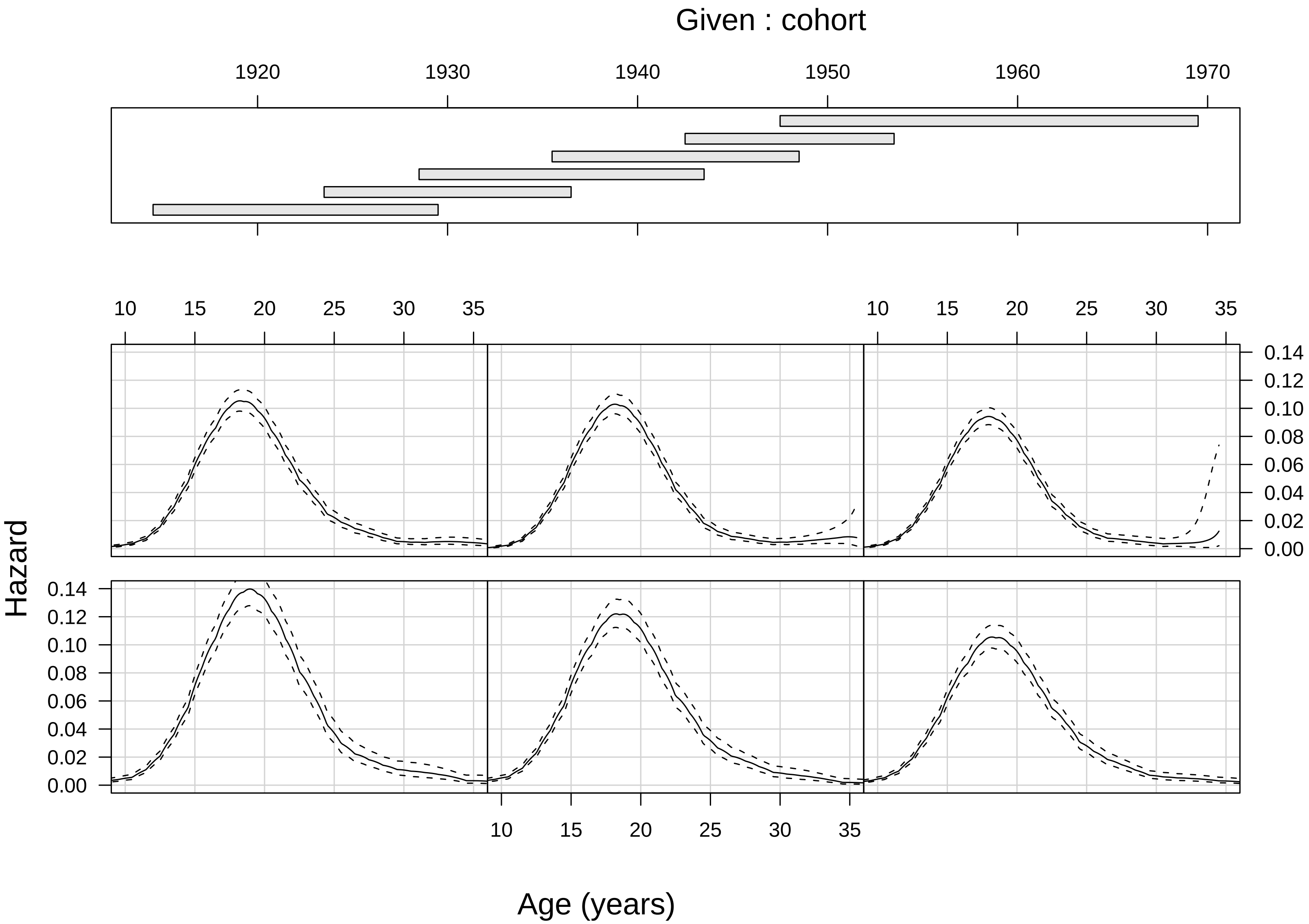


Figure 5.4: Coplot of initiation of tobacco smoking for males from the Risk Factor Prevalence Study

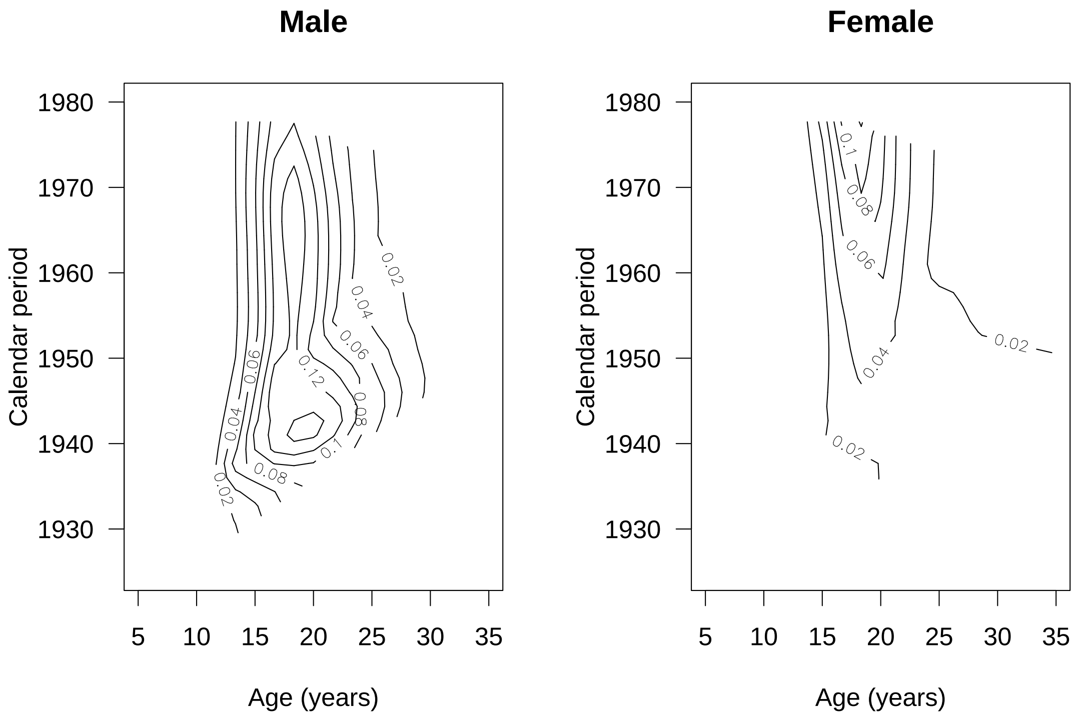


Figure 5.5: Age-period contour plot of the hazard of initiation of tobacco smoking from the Risk Factor Prevalence Study, by sex

penalised splines for cohort, formal tests for proportionality provided strong evidence that both males and females were not proportional ($p < 0.0001$ for both sexes). Restriction to cohorts born before 1930 and 1950 likewise provided evidence that the hazards were not proportional. The following tests for trends should thus be considered exploratory.

Compared with male birth cohorts before 1950, initiation rates for more recent birth cohorts have slowly dropped ($\mathrm{RR} - 1 = -0.012$, 95% confidence interval: -0.016, -0.008).

Compared with female initiation for cohorts born before 1950, there was evidence for a slow linear rise for more recent cohorts ($\mathrm{RR}-1 = 0.011$, 95% confidence interval: 0.005, 0.016). However there was some evidence for convex curvature ($p = 0.07$ for quadratic versus linear model) which suggests that the initiation rates for *older* cohorts may be more stable between cohorts.

5.3.2 Smoking cessation from the Risk Factor Prevalence Study

5.3.2.1 Cumulative incidence

Estimates of cumulative incidence of smoking cessation are shown in Figure 5.6. There were fewer differences between males and females for smoking cessation compared with smoking initiation. In general, estimates of smoking cessation were higher among the younger birth cohorts. The youngest two birth cohorts had greater cessation at the youngest ages, which may reflect greater recall of smoking experimentation by the younger compared to the older cohorts. For the older birth cohorts, smoking cessation under age 20 years was near negligible, as evident by the flat slope of the line.

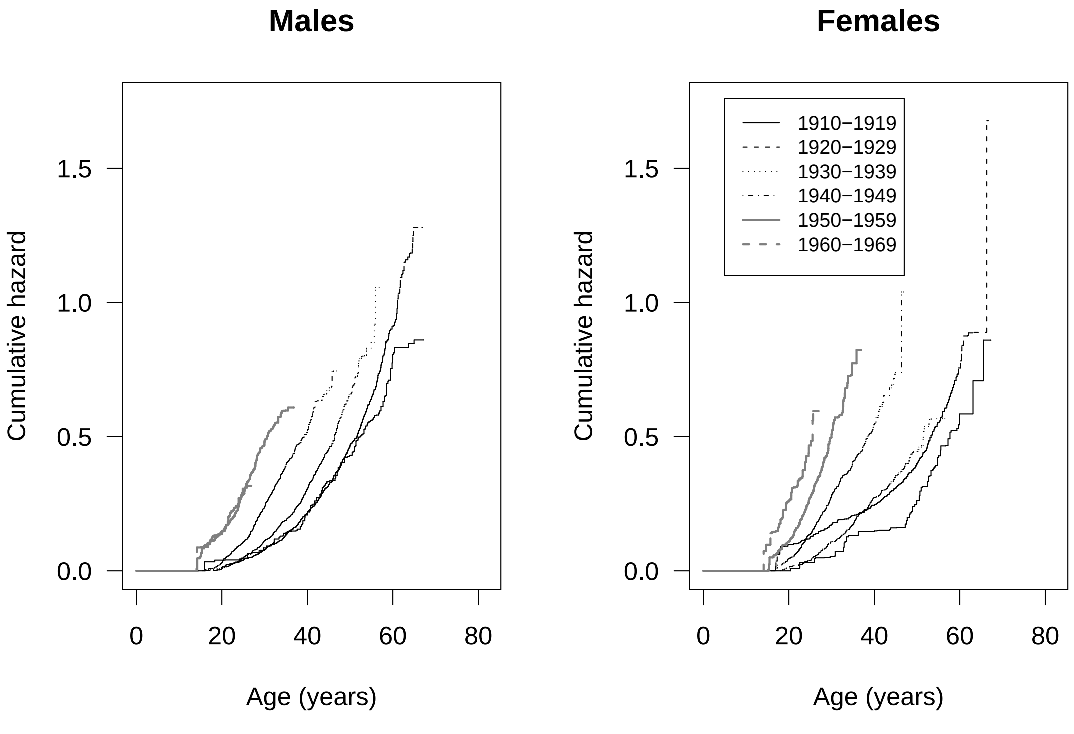


Figure 5.6: Cumulative hazard of cessation of tobacco smoking from the Risk Factor Prevalence Study, by sex

Another general pattern was that the rate of smoking cessation, represented by the *slope* of the cumulative hazard curve, increased with age. The cumulative hazard curves were more consistent for males than for females. The shape and shift of the curves for males were similar between birth cohorts. The lower consistency for females is possibly explained by fewer numbers of smokers, as evident from the cumulative measures of initiation. The inconsistency may make the interpretation of female estimates on the Lexis diagram more difficult.

For females, cessation among the younger cohorts may be higher than the older cohorts. This would be consistent with recent trends for smoking uptake by young females followed by increased cessation during the twenties and thirties.

5.3.2.2 Hazard estimates

From the conditional plot of cessation rates for males (Figure 5.7 on page 134), cessation rates tended to rise linearly between 20 and 50 years of age, with a more rapid increase at older ages. The younger cohorts tended to have higher cessation rates for the same age. There is also a suggestion that the youngest cohorts had higher rates at the youngest ages. Estimates close to the time of the survey were imprecise.

The pattern for female cessation rates was broadly similar to that for males (Figure 5.8 on page 135). The main difference is a more flat pattern by age for females compared with the marginal increase by age for males.

The confidence intervals were very wide at the youngest and oldest ages and the lower bound was sometimes informative. For example, the steep incline after 55 years of age provided a lower bound that was at or above the level for previous ages. As noted in the methods, the confidence interval is only informative for the *smoothed* estimator, which may itself be biased.

For the contour plots of smoking cessation rates on the Lexis diagram, rates tend to rise with period, with early and late peaks for age (see Figure 5.9). The late peak for males was more pronounced, although the pattern at the boundary for the oldest cohorts is unclear. A likely explanation, suggested by the conditional plots, is that the cessation estimates close to the boundaries were imprecise. The range of cessation estimates available by age and period were constrained to those birth cohorts surveyed, so that limited information was available for earlier periods.

5.3.2.3 Proportional hazards models

The most recent birth cohorts had substantially different cessation patterns to those born in earlier years. As expected, there was strong evidence for a departure from proportionality of the hazards ($p < 0.0001$ for both sexes). In a manner similar to initiation, there was evidence for non-proportionality of the cessation hazards for those born before 1950. Again, the trends were estimated for exploratory purposes.

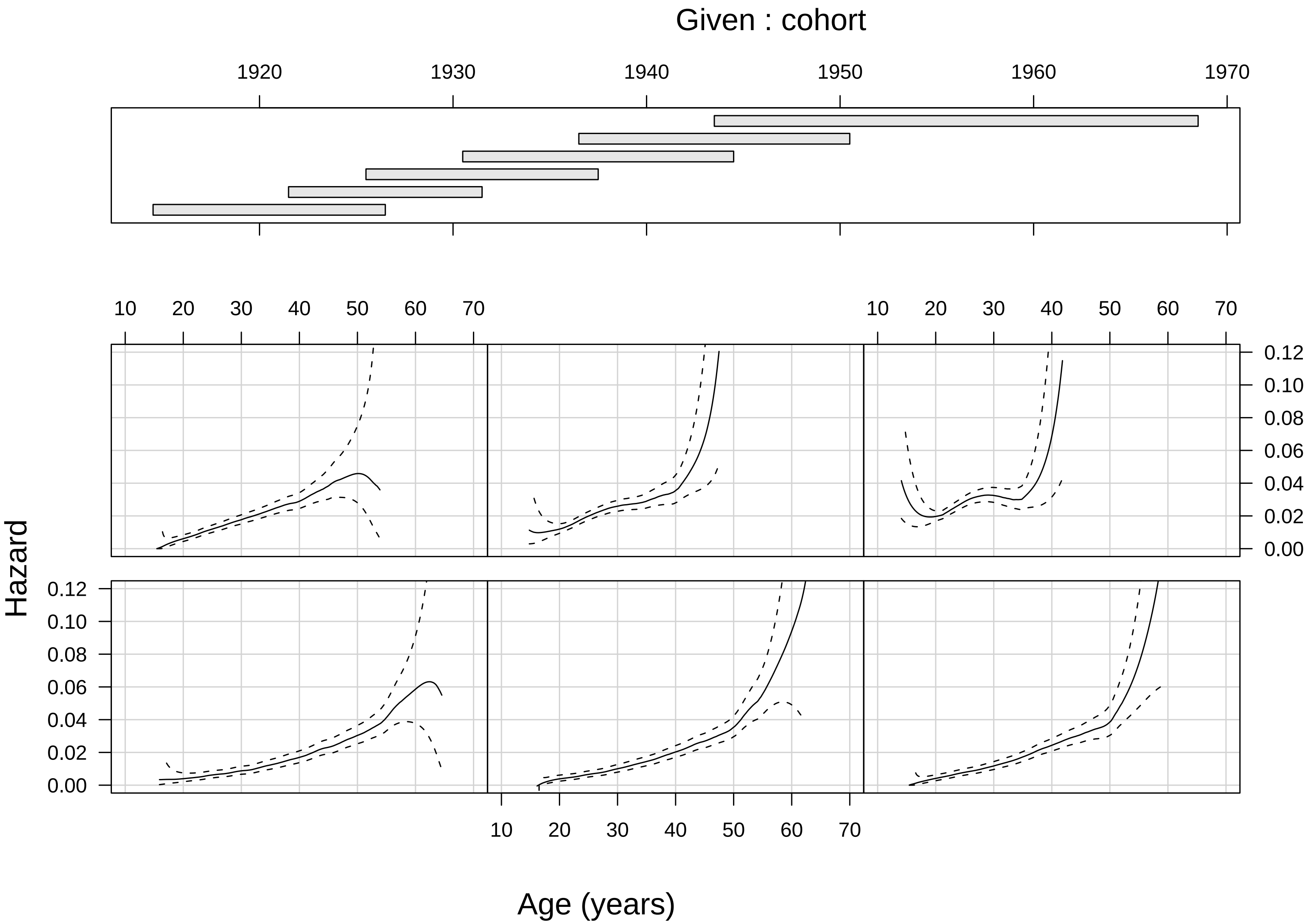


Figure 5.7: Coplot of cessation of tobacco smoking for males from the Risk Factor Prevalence Study

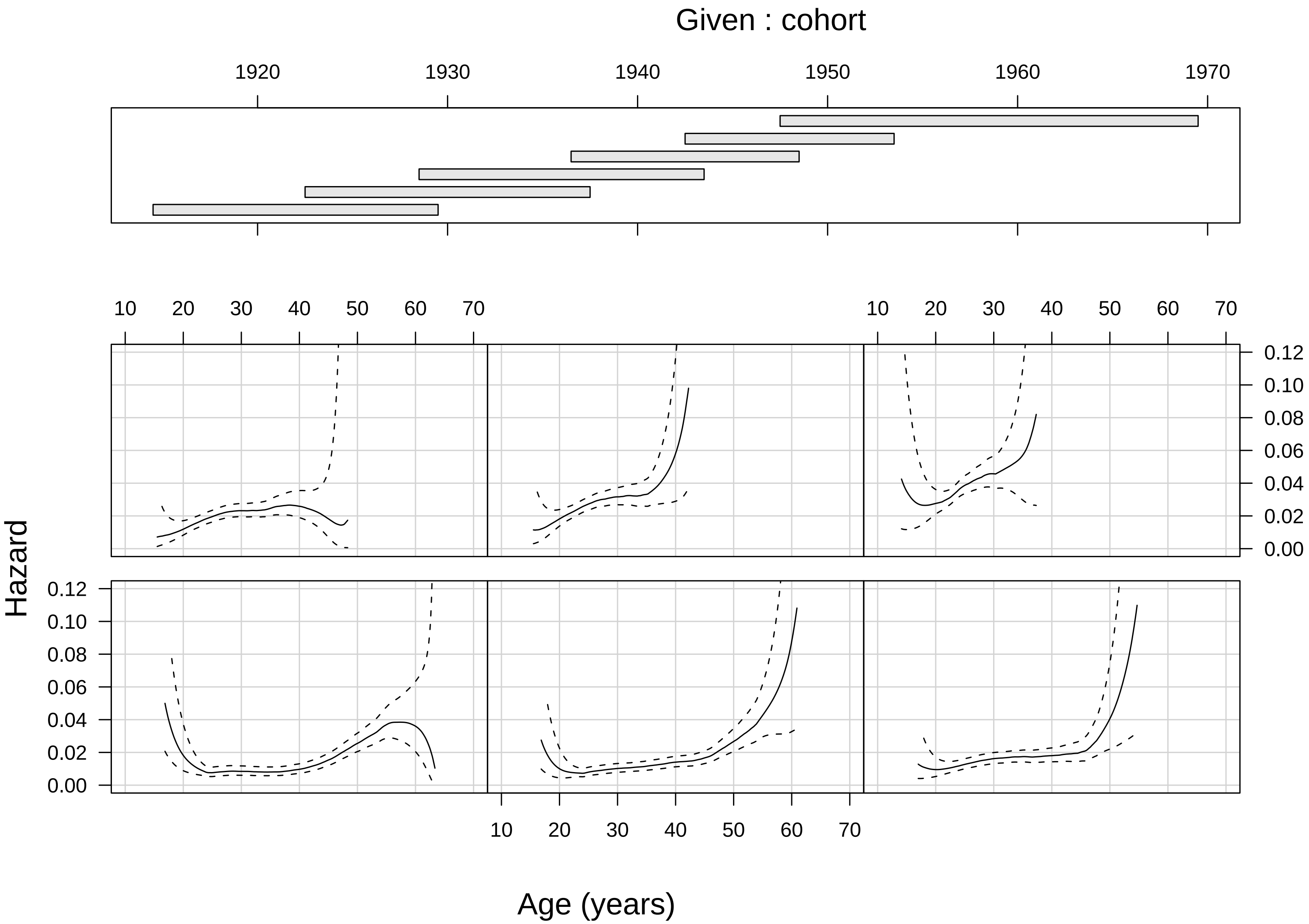


Figure 5.8: Coplot of cessation of tobacco smoking for females from the Risk Factor Prevalence Study

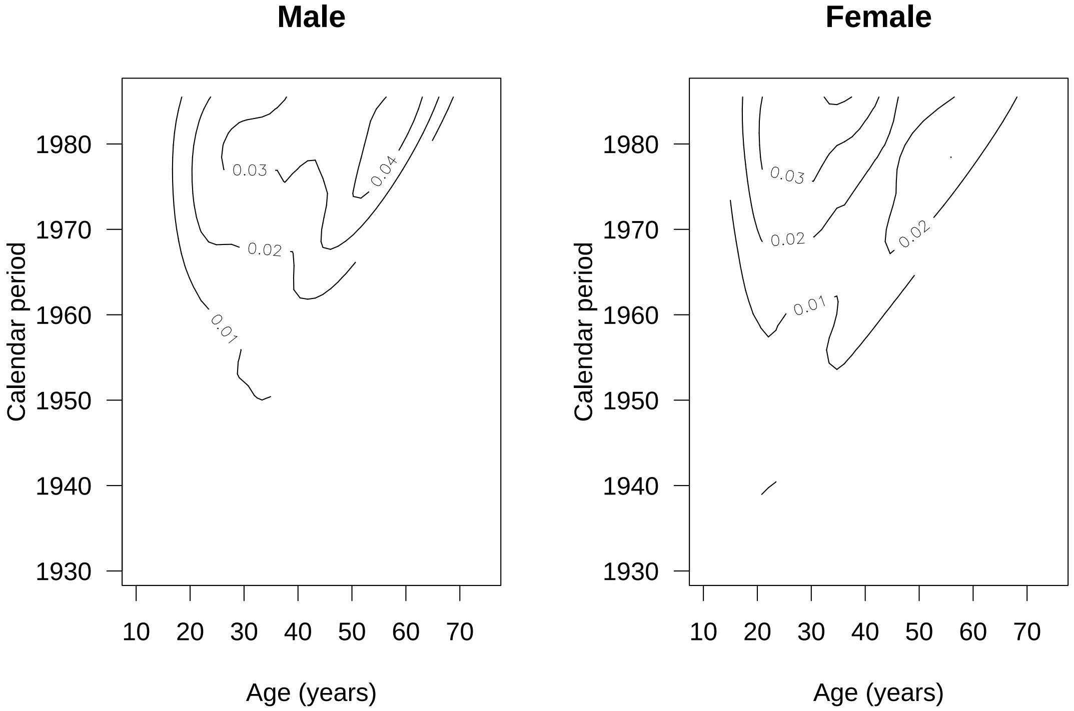


Figure 5.9: Age-period contour plot of the hazard of cessation of tobacco smoking from the Risk Factor Prevalence Study, by sex (data are only available at the top-left of the plot)

For changes in the age-specific cessation rate, the male pattern can be summarised as being piecewise linear, where the increase in cessation rates for more recent cohorts was less for those born before 1935 ($\text{RR} - 1 = 0.033$, 95% confidence interval: 0.016, 0.050) compared with those born during 1935–1949 ($\text{RR} - 1 = 0.072$, 95% confidence interval: 0.051, 0.092). Females followed a similar pattern, with the rise for birth cohorts born before 1935 ($\text{RR} - 1 = 0.033$, 95% confidence interval: 0.009, 0.059) being less than half of the increase compared with those born during 1935–1949 ($\text{RR} - 1 = 0.082$, 95% confidence interval: 0.054, 0.109).

5.3.3 1998 National Drugs Strategy Household Survey

5.3.3.1 Introduction

The 1998 National Drugs Strategy Household Survey has more recent data and collected information from a broader age range than the Risk Factor Prevalence Study.

However, the Risk Factor Prevalence Study has several advantages. First, the sample size for the Risk Factor Prevalence Study is considerably larger than that collected for the 1998 National Drugs Strategy Household Survey (22,566 v. 10,030). Second, the National Drugs Strategy Household Survey had three different sampling frames, of which two sampling frames were based on self-completion of questionnaires collected in different manners. Ideally, useful data would be restricted to the sampling frame where the data were collected face-to-face, however, this would further constrain the sample size. The alternative is to include all respondents, which may affect validity. Third, although the Risk Factor Prevalence Study collected data for a restricted age range, being less than 65 years for the first two surveys and less than 70 for the last survey, the birth cohorts represented are comparatively old compared with data collected by the 1998 National Drugs Strategy Household Survey. Moreover, the sampling frames used by the 1998 National Drugs Strategy Household Survey were designed to over-sample for younger people, so that there are few data for older respondents.

Although the National Drugs Strategy Household Survey does provide detail for cohorts born after 1960, the validity of retrospective data for recent or future rates is of concern. Given the evidence in Chapter 6 that retrospective initiation estimates may be imprecise or biased, the preference is to use current status data to predict the more recent rates.

This suggests that estimates from the 1998 National Drugs Strategy Household Survey should be compared with estimates from the Risk Factor Prevalence Study for the same time period. However this would be open to interpretation, as there could be variation between the two data sources due to differential recall dependent on time since a respondent was interviewed.

In that context, the results from the National Drugs Strategy Household Survey are provided here for general comparisons, although specific comparisons with the 1996/97 New Zealand Health Survey may be more informative and reliable.

5.3.3.2 Comparisons

From the Lexis diagrams, the two Australian data sources exhibit broadly similar contemporaneous patterns for estimates of both initiation and cessation (comparing Figures 5.5 with 5.13, and comparing Figures 5.9 and 5.17). Given the potential limitations discussed earlier in this section, this suggests that the method may be robust.

More recent initiation estimates suggest that male rates are either stable or falling (see Figures 5.10–5.12). There is no evidence from these data that female initiation rates have peaked.

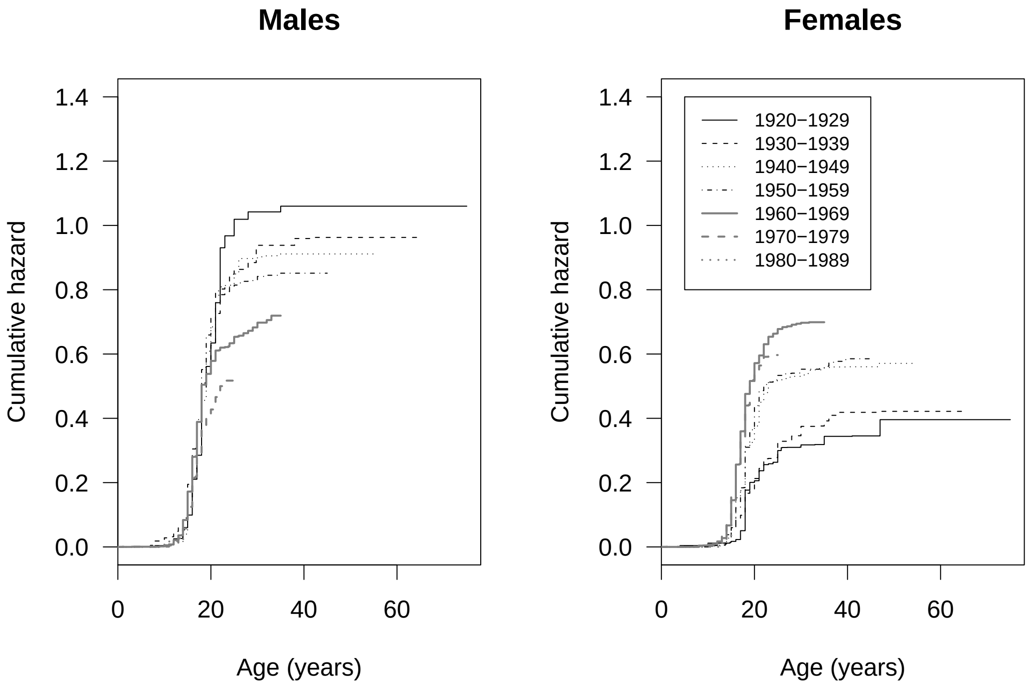


Figure 5.10: Cumulative hazard of initiation of tobacco smoking from the 1998 National Drugs Strategy Household Survey, by sex

For younger birth cohorts male cessation rates look relatively flat across ages, at a level of 0.02 to 0.025 per year (see Figures 5.14–5.15). The cessation rates among younger females have a distinct curvature by age, peaking at around 0.04 per year between ages 25–30 years (see also Figure 5.16).

5.3.4 1996/97 New Zealand Health Survey

The sample size for the 1996/97 New Zealand Health Survey ($n = 7755$) was considerably smaller than that for the Risk Factor Prevalence Study ($n = 22,566$). Consequently, the confidence intervals were not as tight; this was most noticeable for the oldest age groups. Interestingly, the results for the two Australian data sources and the 1996/96 New Zealand Health Survey were quite similar.

The most recent female cohort initiation rates from the 1996/97 New Zealand Health Survey were lower than the previous birth cohorts (see Figures 5.18–5.20). This finding is different from the pattern for the two youngest birth cohorts from the National Drugs Strategy Household Survey, where no change was apparent, and from the Risk Factor Prevalence Study, where the rates increased for the younger birth cohorts. The pattern of decreasing male initiation for younger birth cohorts

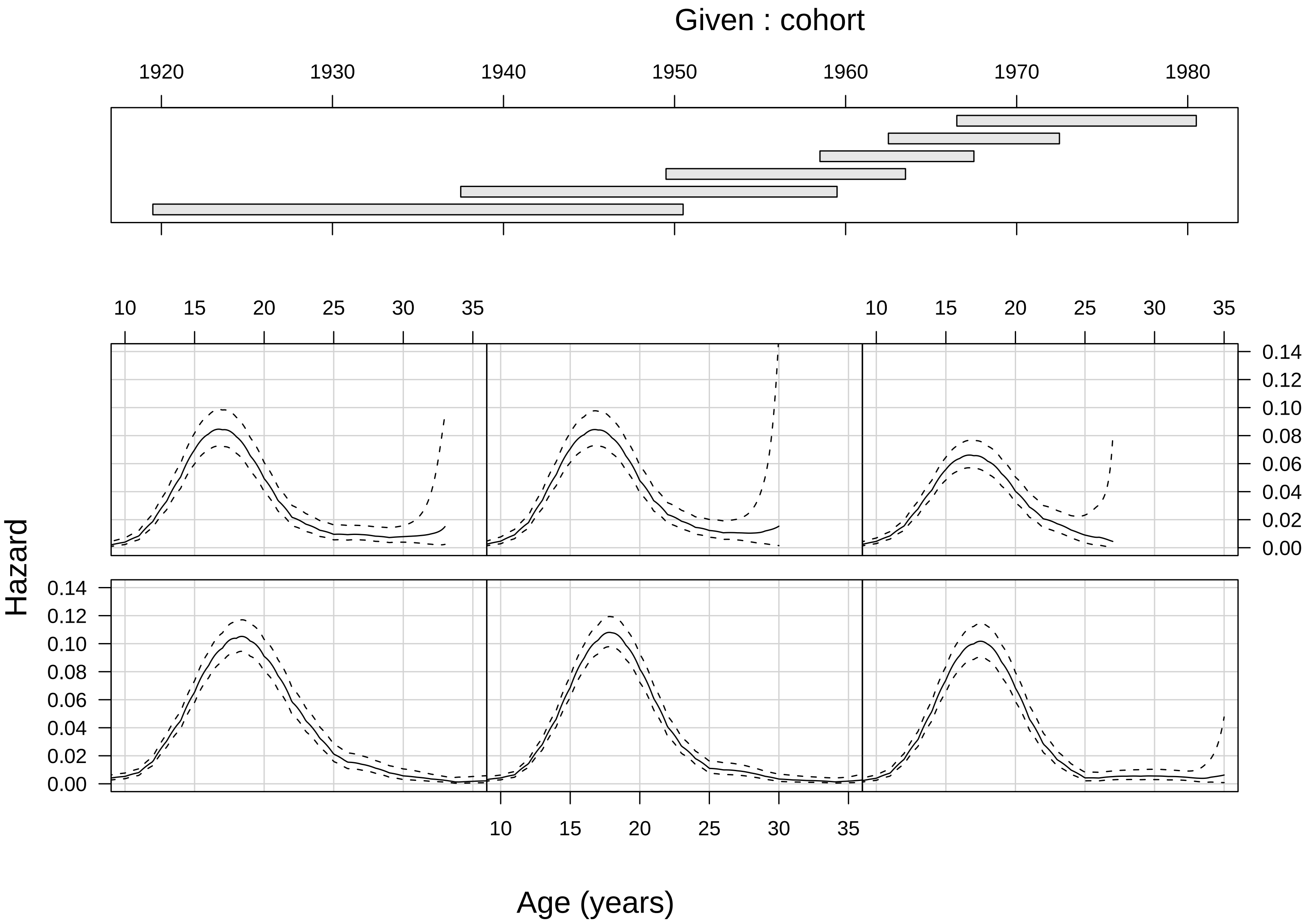


Figure 5.11: Coplot of initiation of tobacco smoking for males from the 1998 National Drug Stategy Household Survey

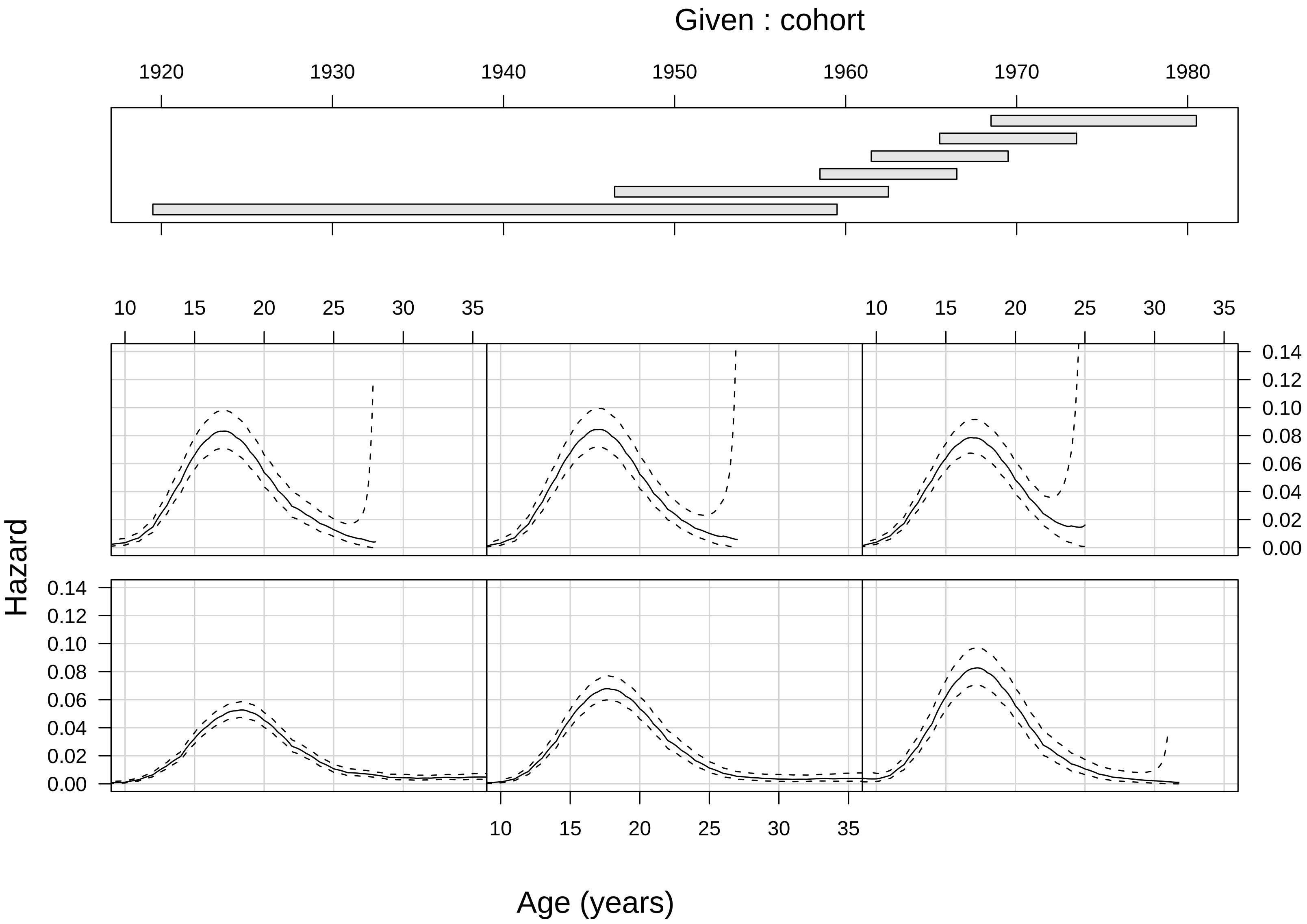


Figure 5.12: Coplot of initiation of tobacco smoking for females from the 1998 National Drug Stategy Household Survey

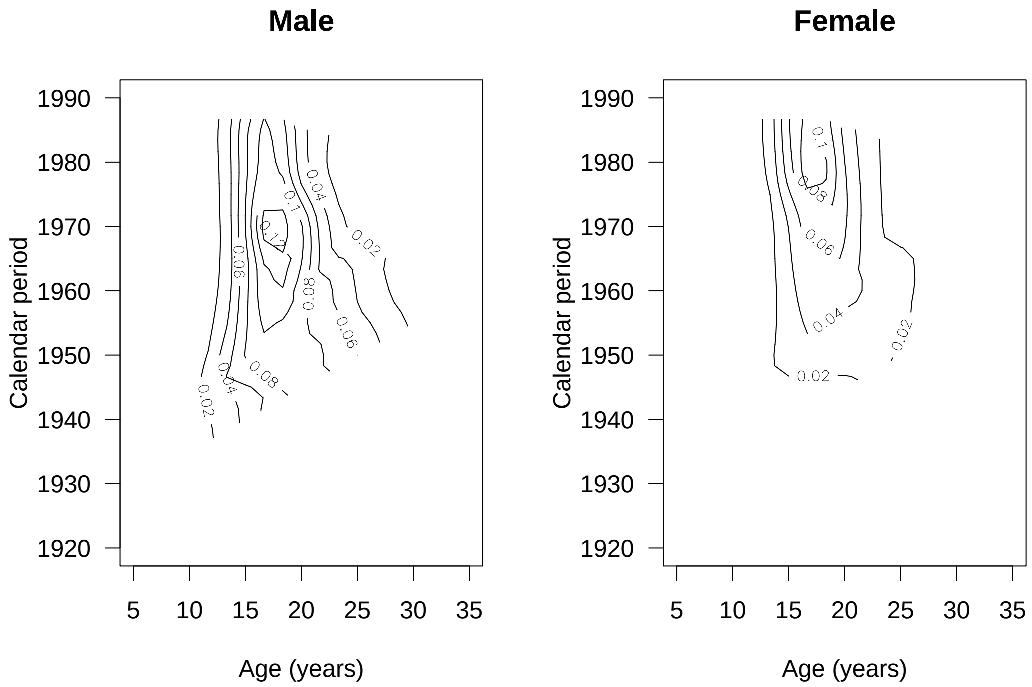


Figure 5.13: Age-period contour plot of the hazard of initiation of tobacco smoking from the 1998 National Drug Strategy Household Survey, by sex

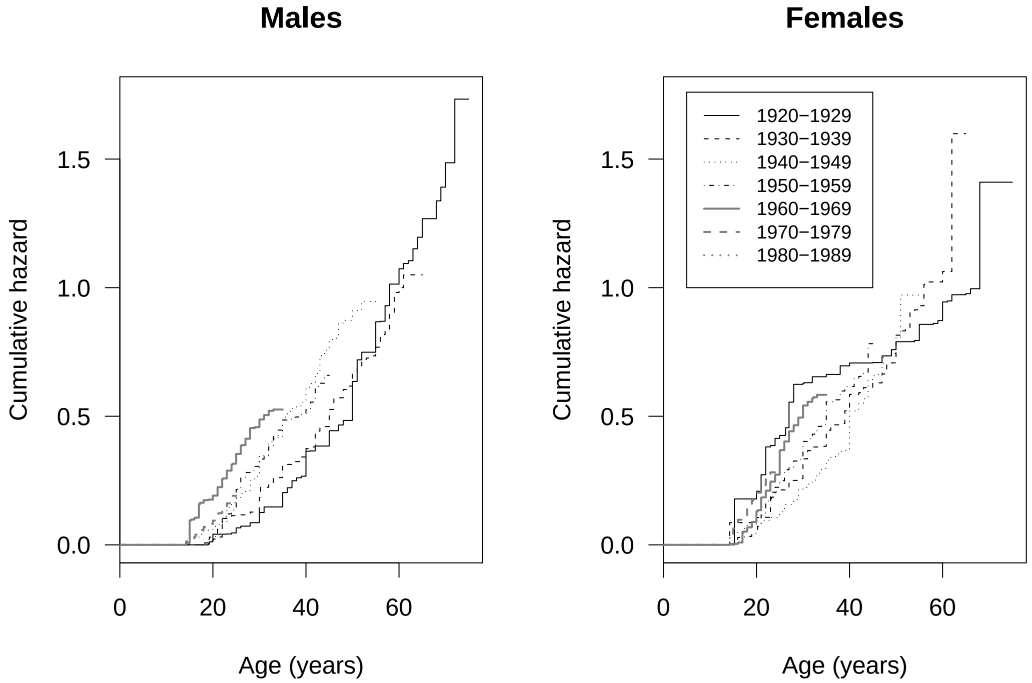


Figure 5.14: Cumulative hazard of cessation of tobacco smoking from the 1998 National Drugs Strategy Household Survey, by sex

was consistent between the three data sources.

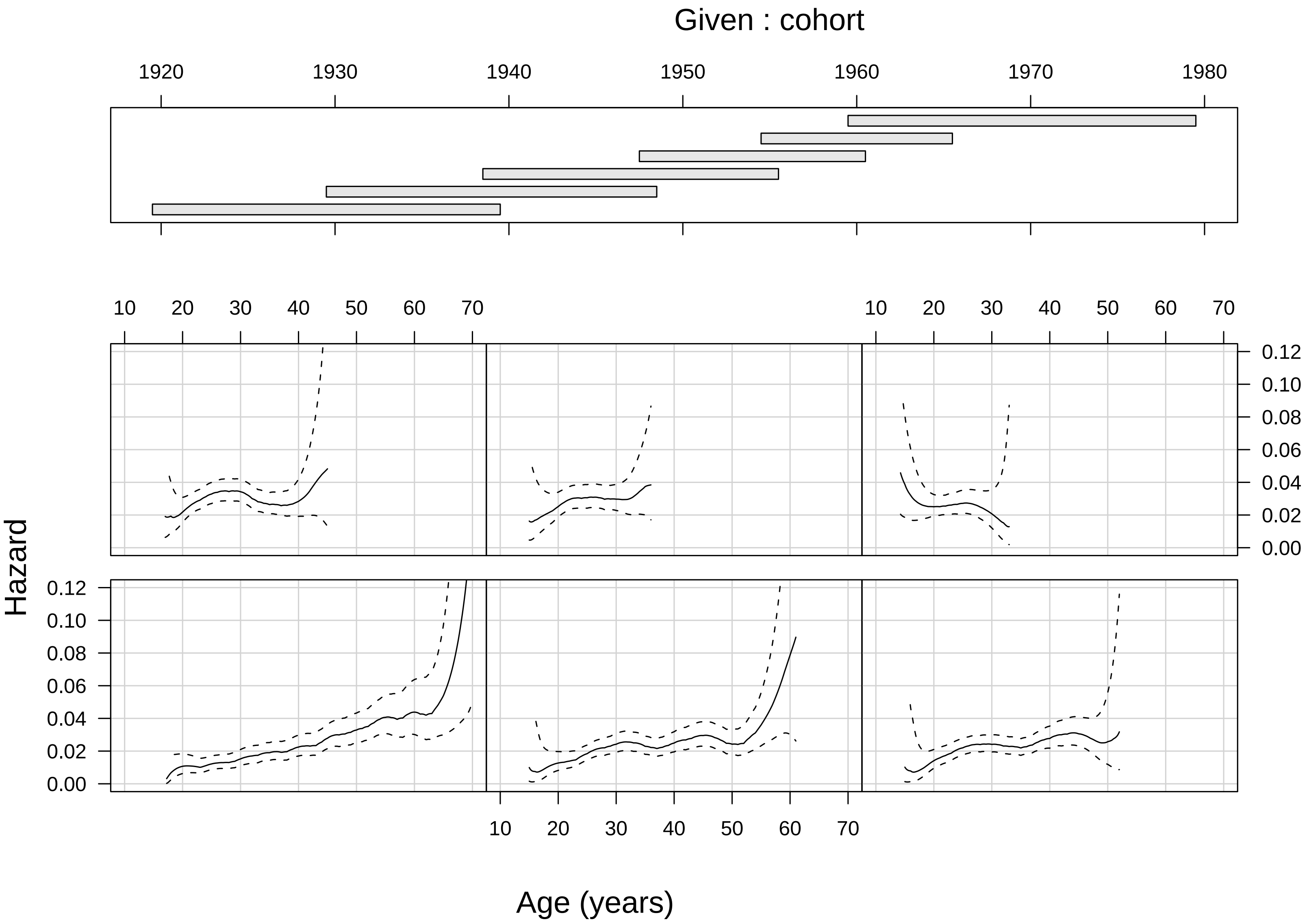


Figure 5.15: Coplot of cessation of tobacco smoking for males from the 1998 National Drug Stategy Household Survey

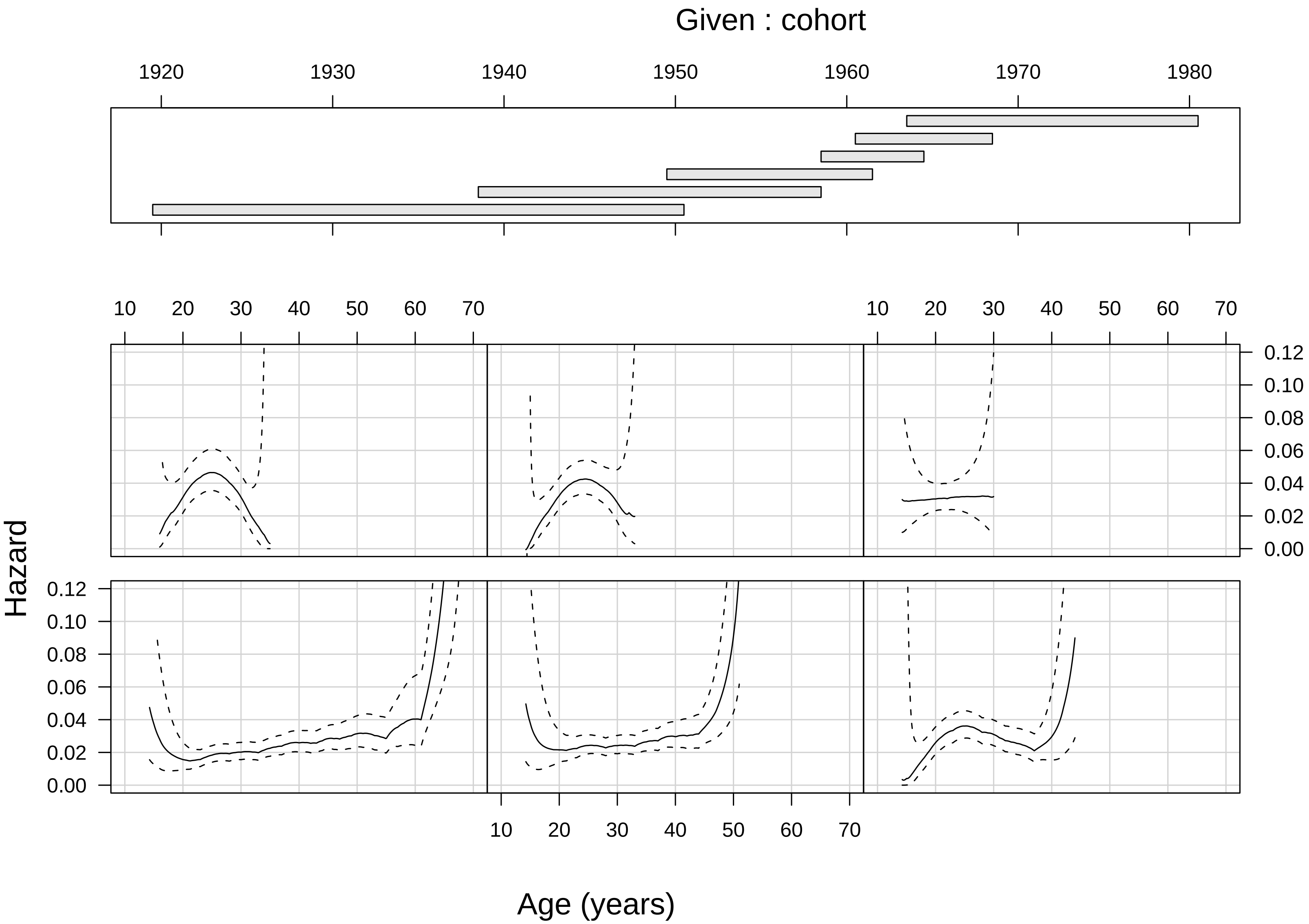


Figure 5.16: Coplot of cessation of tobacco smoking for females from the 1998 National Drug Stategy Household Survey

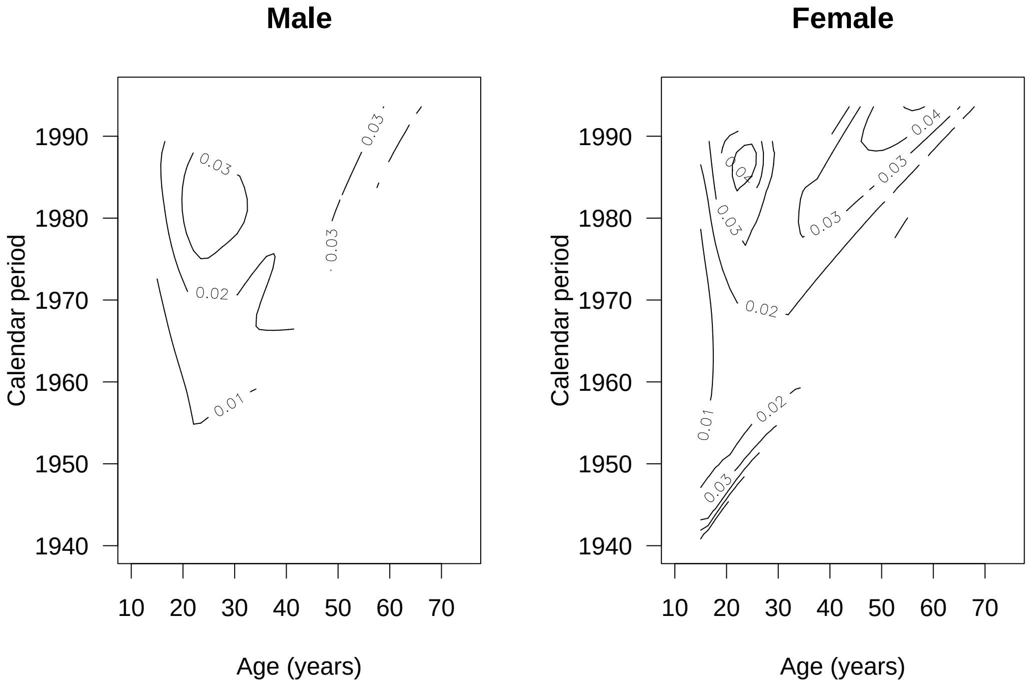


Figure 5.17: Age-period contour plot of the hazard of cessation of tobacco smoking from the 1998 National Drug Strategy Household Survey, by sex

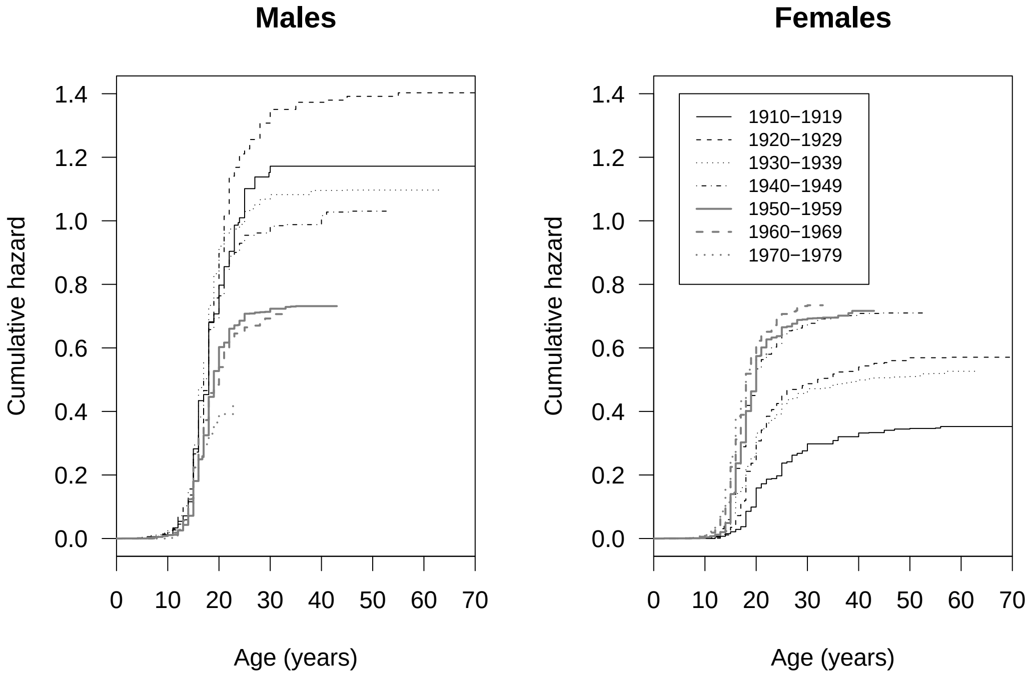


Figure 5.18: Cumulative hazard of initiation of tobacco smoking from the 1996/97 New Zealand Health Survey, by sex

The conditional plots for cessation rates were remarkably similar to those from the Risk Factor Prevalence Study (Figures 5.22–5.20). Cessation rates among the

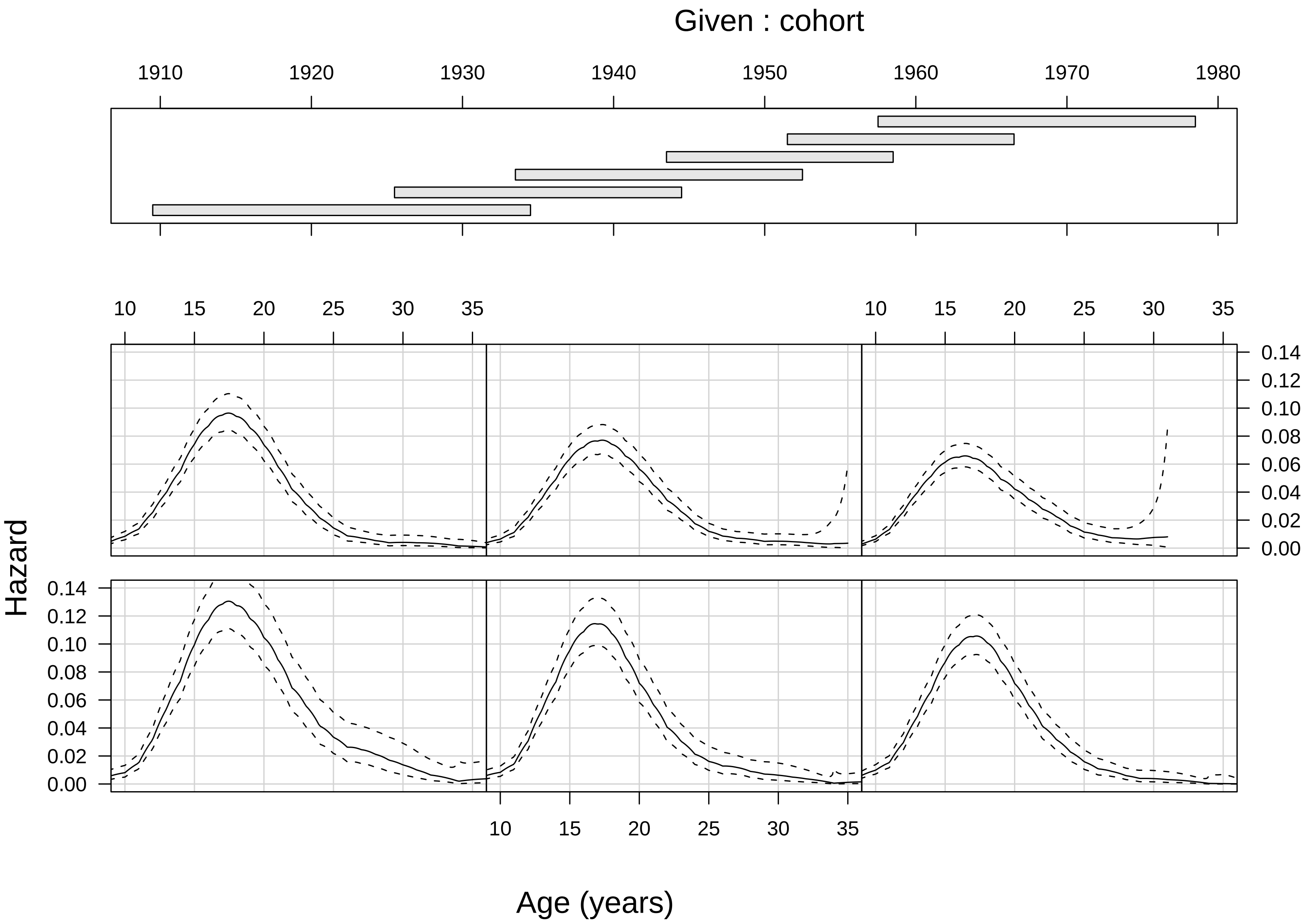


Figure 5.19: Coplot of initiation of tobacco smoking for males from the 1996/97 New Zealand Health Survey

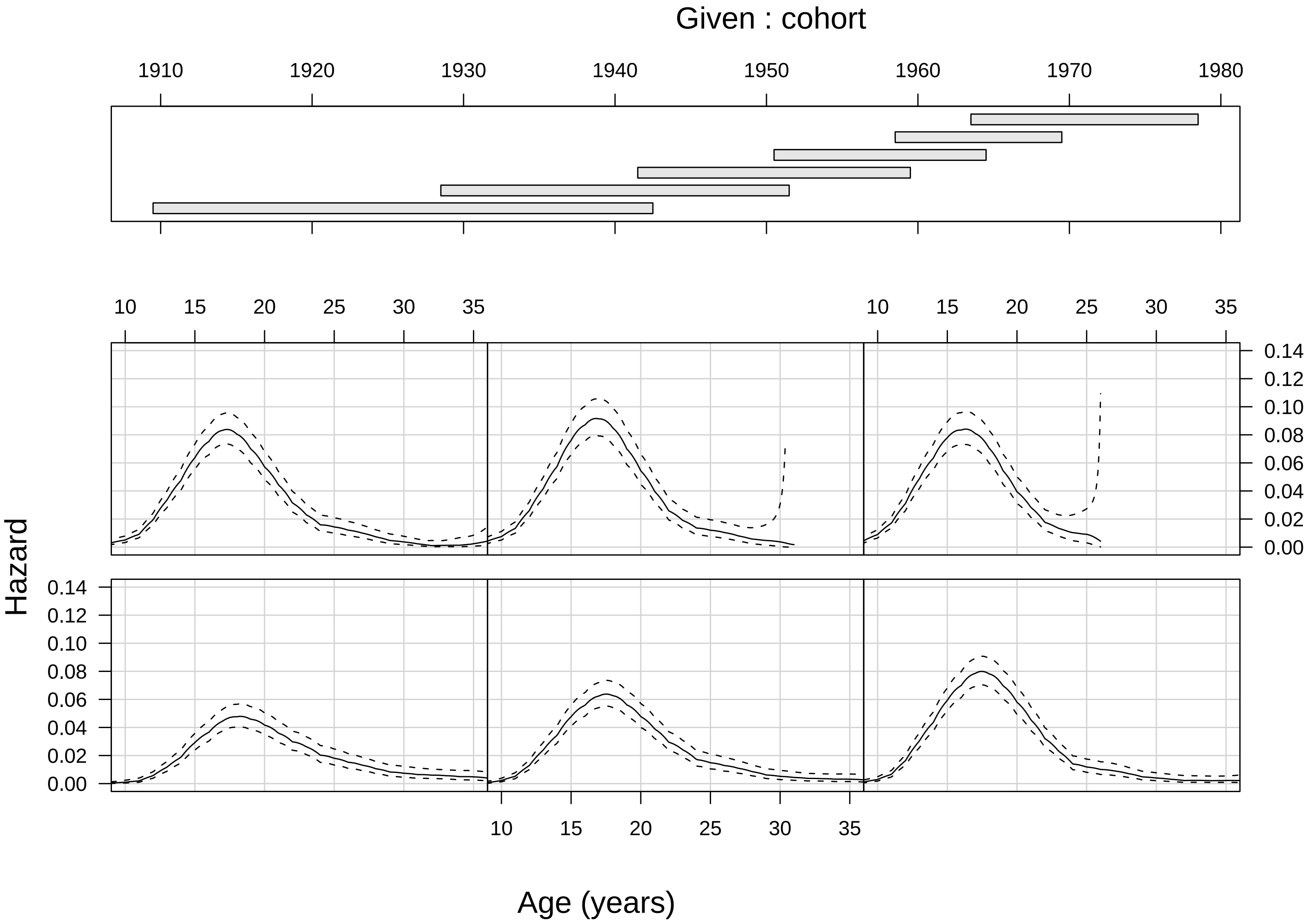


Figure 5.20: Coplot of initiation of tobacco smoking for females from the 1996/97 New Zealand Health Survey

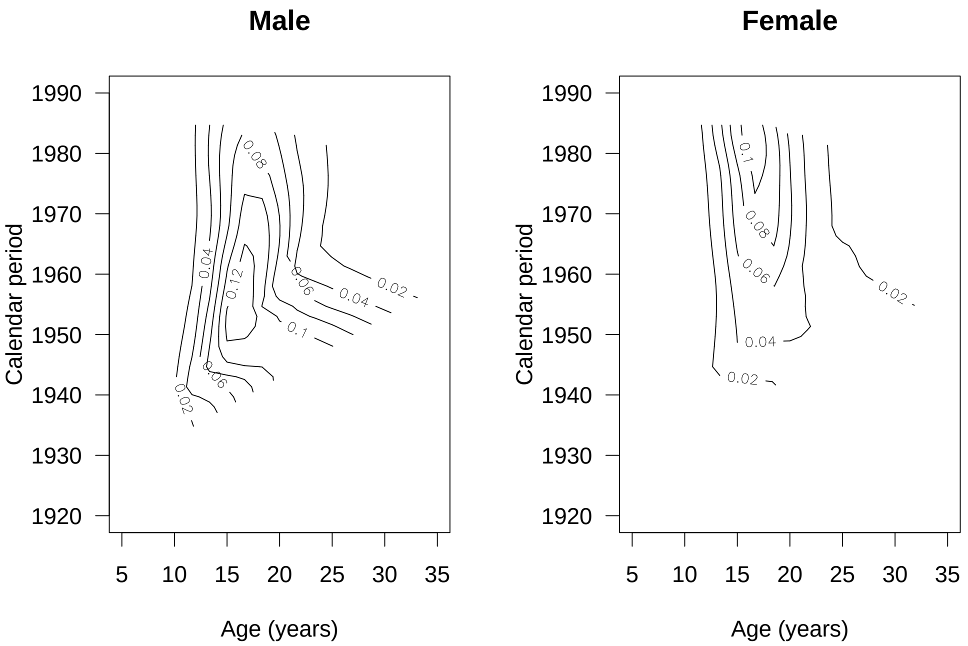


Figure 5.21: Age-period contour plot of the hazard of initiation of tobacco smoking from the 1996/97 New Zealand National Health Survey, by sex

youngest female cohorts were concave by age, peaking at ages 25–30 years, which is consistent with estimates from the National Drugs Strategy Household Survey. Moreover, there was little evidence for changes in cessation rates between the younger birth cohorts for the two surveys.

Similar to the Risk Factor Prevalence Study, there was evidence against proportionality of the hazards for either initiation or cessation.

For males born before 1950, there was no evidence for any significant change in initiation rates across birth cohorts ($\mathrm{RR} - 1 = -0.005$, 95% confidence interval: -0.014, 0.003), although the spline fit suggested that rates may have peaked around 1920–1930. Female initiation rates for cohorts born before 1950 were lowest for the oldest cohorts, rising over time ($\mathrm{RR} - 1 = 0.020$, 95% confidence interval: 0.010, 0.030).

For cohorts born before 1950, males experienced a linear increase in cessation rates ($\mathrm{RR} - 1 = 0.029$, 95% confidence interval: 0.016, 0.043). In contrast, the female cessation rates were at a minimum during the late 1920s, rising rapidly for the younger birth cohorts (see Figures 5.22– 5.24).

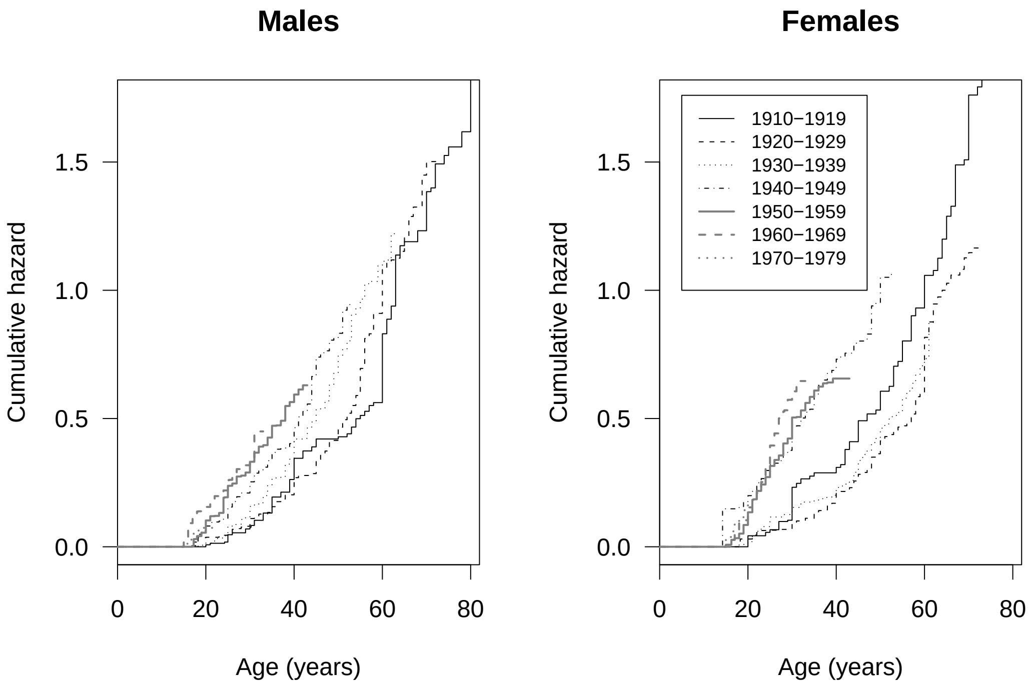


Figure 5.22: Cumulative hazard of cessation of tobacco smoking from the 1996/97 New Zealand Health Survey, by sex

5.4 Discussion

5.4.1 Summary

For Australian birth cohorts born during 1910–1969, rate estimates of smoking initiation among males declined from the older to the more recent birth cohorts. In contrast, female rates were lowest for the oldest birth cohorts and increased across the younger birth cohorts, with a 5–10% increase in peak prevalence of ever smoking.

Cessation rates rose at older ages and for more recent periods. For the most recent periods, there was a suggestion of more rapid cessation at younger ages, which may be due to recall of experimentation. For male cohorts born before 1910, the predicted trends suggest that cessation rates may have been at similar or higher levels compared with younger cohorts, while rates for older female cohorts may have been less than those for the younger female cohorts.

5.4.2 Limitations

There are several limitations to this study. First, data were only available for birth cohorts born during 1910–1969. Older birth cohorts would have provided important information about earlier trends in initiation and cessation. The description of initi-

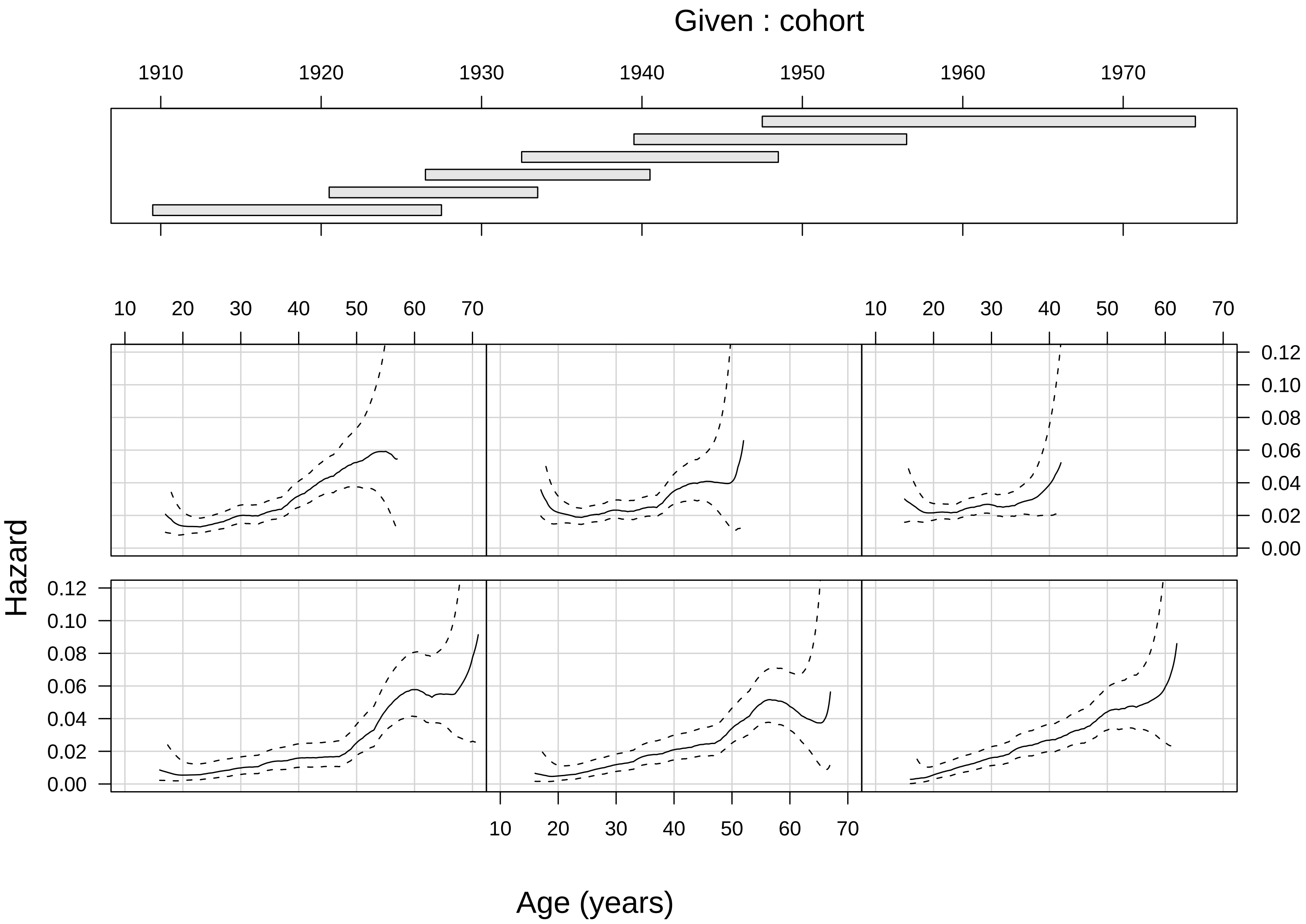


Figure 5.23: Coplot of cessation of tobacco smoking for males from the 1996/97 New Zealand Health Survey

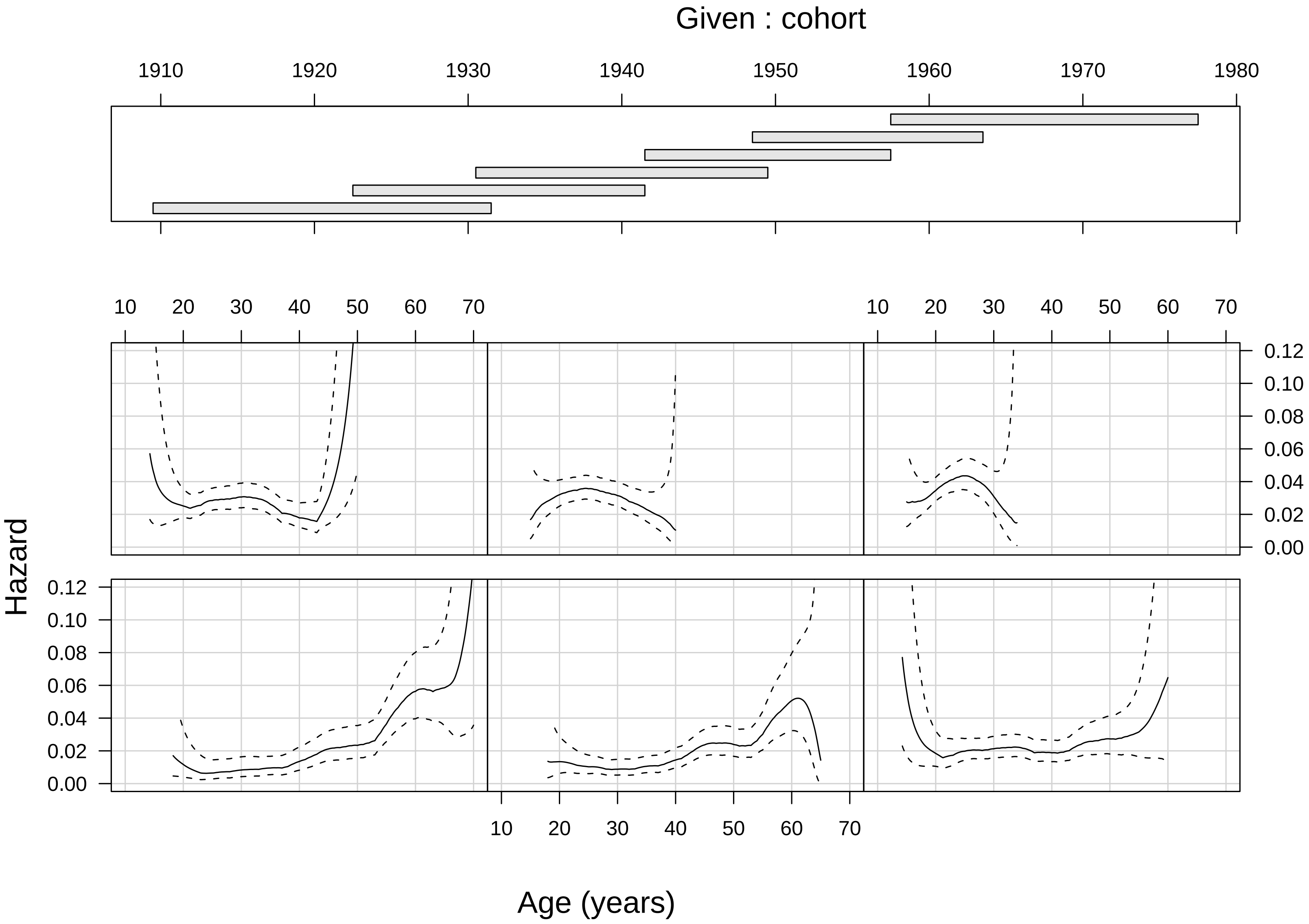


Figure 5.24: Coplot of cessation of tobacco smoking for females from the 1996/97 New Zealand Health Survey

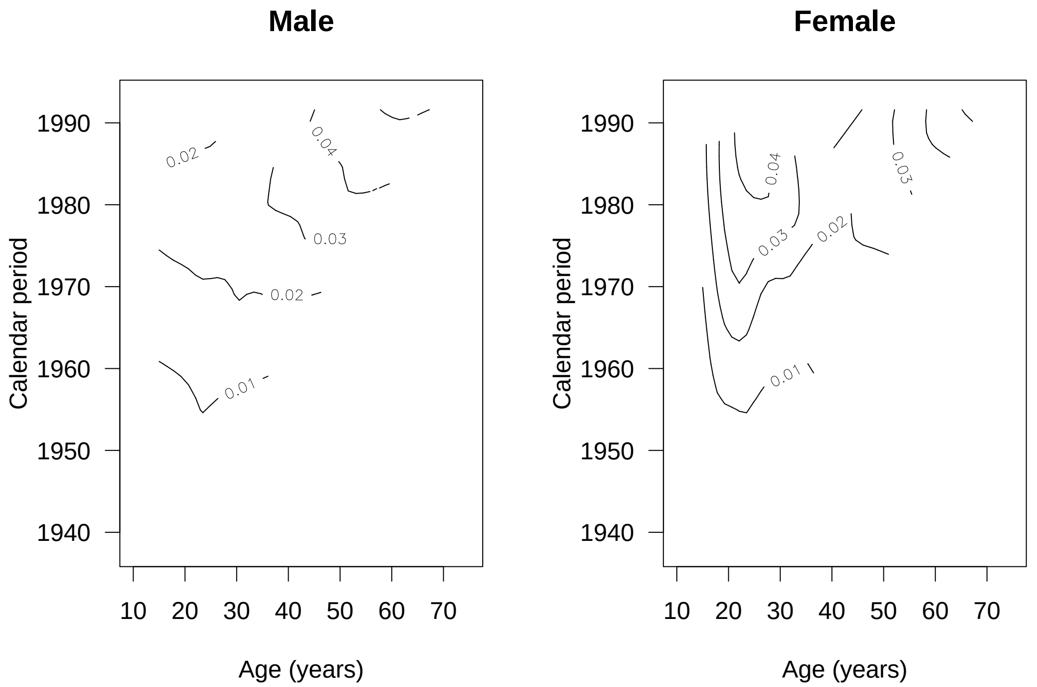


Figure 5.25: Age-period contour plot of the hazard of cessation of tobacco smoking from the 1996/97 New Zealand National Health Survey, by sex

ation is more certain than cessation because initiation is largely limited to younger ages.

Second, it is more difficult to precisely measure rates of change of prevalence than prevalence itself. Although the Risk Factor Prevalence Study was of a moderate size ($n = 22,566$), precise estimates were not available by five-year age groups.

Third, cessation estimation is imprecise at the boundaries of the data on the Lexis diagram. In particular, estimates for the oldest birth cohorts were too imprecise to be useful. One consequence is that greater care is required for projections outside of the boundaries.

Fourth, the validity of retrospective estimates requires further assessment. The rise in female cessation rates for the period immediately prior to interview suggests systematic bias due to recall of recent events. Consequently, prediction of cessation rates for most recent years and projections for later years may have to rely on current status data. Projections are particularly sensitive to model assumptions and data quality, both of which may affect validity (see Mendez et al., 1998, and Appendix B). Survey data prior to the 1970s were not available, so that the retrospective estimates need to be used and can only be validated using similar such data sources. Estimates based on retrospective and current status data can be expected to give qualitatively different results, as experimentation would be more easily estimated from current status data. Similarly, short-term smoking is likely to be poorly recalled for older cohorts. Finally, given data recalled at one point in time, biases due to period or cohort effects are totally confounded and cannot be assessed.

5.4.3 Methodological implications

With regards to methodological implications, the choice of using cumulative hazards or survival for presentation was equivocal. The survival estimate for initiation, being in this case (one less) the prevalence of ever smokers, has a simple lay interpretation. The interpretation of survival estimates for cessation is more complicated because of uptake at different ages. In contrast, the slope of the cumulative hazard is interpretable as the hazard for both initiation and cessation, which may explain why the cumulative hazard is in common use by statisticians. The cumulative hazard was chosen because of its consistent interpretation, while survival would have been chosen had attention been restricted to initiation.

The adjustment for differential mortality only affected the estimates for the oldest birth cohorts, as differential survival is limited under a nominal age of 50 years.

5.4.3.1 Computational approaches

The computational approaches chosen here were based on providing valid estimates using the simplest possible methods. The non-parametric estimators provided a description of smoking initiation and cessation and suggest alternative estimation strategies. A straightforward approach would be to fit parametric hazard models using parametric forms as suggested by the results in this chapter. Parametric models for left truncated and right censored observations are available in `S-Plus` and in `R` (from July 2001) using the `gss` library.

An alternative strategy would be to assume smooth hazard functions using splines, fitting the hazard with possible covariates using penalised likelihood. Joly et al. (1999) have provided estimation software (`PHMPL`) for unweighted observations that are possibly right censored and left truncated or interval censored. Unweighted analyses using `Fortran` code for `PHMPL` interfaced with `R` software by the author suggests this would be a flexible approach. This approach has not been reported in detail as weighting for the study design was considered an important aspect of the analysis. Useful software developments could include the incorporation of weights in `PHMPL` code and the improvement of an `R` interface.

In this chapter local likelihood hazard estimation was performed using functions for density estimation. Cohort-specific hazard curves using this approach were similar to those using the kernel-based approach. Software for local likelihood estimation of the hazard directly incorporating weighted left truncated and right censored data, tools for which are not known to be currently available, would be valuable.

Another strategy related to the issue of hazard estimation on the Lexis diagram has been outlined in a technical article by Ogata et al. (2000), where empirical Bayesian models are used for spline fitting, including facilities for model comparison.

Hazard estimation continues to be an active area of statistical research. One difficulty for the applied statistician is the availability of software to perform analyses that are considered established in the statistics literature. One outcome of the chapter is the development of simple tools based around existing software to perform novel tasks. With some effort, the available analytic tools can be readily extended, however it is important that practitioners understand how to use them.

5.4.4 Interpretation

An important outcome of the comparisons between the Risk Factor Prevalence Study, the 1998 Australian National Drugs Strategy Household Survey and the 1996/97 New Zealand Health Survey is that estimates were consistent between data sources. This provides a source of validation for both the Risk Factor Prevalence Study and the 1998 National Drugs Strategy Household Survey. Moreover, the broad and specific similarities between the Australian results and the 1996/97 New Zealand Health Survey support the notion that secular trends in Australia and New Zealand have been similar (see also Chapter 1). Consequently, results that are not available from one country, such as estimates for rates of recall error, may potentially be valid for the other country.

Although the 1998 National Drugs Strategy Household Survey provides more recent data, the Risk Factor Prevalence Study was chosen as the primary Australian data source. The motivation for this choice includes the larger size of the Risk Factor Prevalence Study and potential design issues with the 1998 National Drugs Strategy Household Survey.

One important consequence is an improved understanding of changes in population smoking behaviour. Earlier work has tended to emphasise the cumulative hazards or survival, which has summarised most of the pattern. However estimation of the hazards themselves revealed the changing form for the initiation hazard and provided a concrete description of the hazard for cessation.

The level of the hazard for initiation and cessation will be affected by recall error. Recall error by former smokers as never smokers is another example of differential selection, similar to differential mortality. Aside from affecting the estimated level, recall error would affect the shape for the initiation and cessation rates if their patterns were differential to those ever smokers who did not change their recall. Given the difficulty in estimating the rate of recall error, estimation of differential hazards of initiation or cessation for those who changed their recall would be a challenging task.

5.4.5 Comparison with US estimates

The following section compares and contrasts some of the US retrospective estimates provided by Burns et al. (1997b), and constructively critiques their methods. Their study included 460,254 respondents, which is considerably larger than any other similar study.

While there were strong similarities between results for the Australian population and patterns found in United States populations (Burns et al., 1997b), several departures are worthy of commentary. First, the peaks of ever smoking prevalence for Australian males and females were lower than those experienced in the United States. Second, the prevalence of ever smoking for the US female birth cohort born in 1960–1969 was marginally higher than older cohorts, while the prevalence for US females born in 1935–1944 were marginally higher than those for younger birth cohorts. Younger women may have higher levels of cessation in recent years. Moreover, their initiation rates were higher at the same time. One possible interpretation is smoking experimentation during that period, including smoking initiation followed by cessation.

The adjustment for differential survival was based on a regression approach for ever smoking after 30 years of age for a given cohort. Unfortunately, the model assumed that initiation and recall error by former smokers was negligible, such that ever smoking without differential mortality is assumed to be close to constant after age 30 years. As observed in Chapter 6, recall error is *not* negligible. Recall error by former smokers will appear to the regression model as differential mortality and is expected to bias ever smoking prevalence. No adjustment was made for differential survival in estimating smoking cessation. More importantly, the approach used for cessation estimation used actuarial methods and it is not clear whether observations were treated as being left truncated.

In summary, these results may indicate that revision of the US estimates is worthwhile, taking account of differential survival and recall error.

5.4.6 Extensions

In the absence of survey data prior to the 1970s, retrospective data provide possibly the best option for estimating changes in smoking behaviour.

An alternative approach would be to make use of tobacco consumption as a measure of smoking uptake and cessation. Consumption is an important and potentially reliable indicator, however a number of assumptions are required for dose, average age of uptake and general pattern of cessation. The latter two parameters could be provided from this analysis. The main difficulty for both approaches is that no information is known, implicitly or explicitly, about the earlier birth cohorts.

Given potential concerns with the retrospective data, the retrospective estimates with census data in the following chapter.

5.A Simulation data used for algorithm validation

The following `R/S-Plus` code was used to generate random data for estimation of the cessation rate (represented by `alpha.Q`). In words, the data would be from a survey held in 1996, with birth cohorts from 1910 to 1979, with $n = 50$ for each birth cohort. Start times were assumed to be normally distributed with mean equal to age 16 years and an arbitrary standard deviation of 1 year. To avoid numerical instability, initiation before age 10 years was assumed at age 10 years. Moreover, initiation less than five years prior to the survey was assumed to be exactly five years prior. Stop times were assumed (simply) to have an exponential distribution from time of initiation, in which case the hazard is constant and the model is both Markov and semi-Markov. Cessation after the age at survey (=`survey.age`) was assumed censored and the event time was transformed to the age at survey. The data assume equally weighted data, with a design effect of 2.00. Differential survival was not included in the simulation data because the intention was to recover the cessation rate without bias.

```
sim.data <- NULL
for (cohort in 1910:1979) {
  survey.age <- 1996 - cohort
  n.rep <- 50
  alpha.Q <- 0.03        # this is what we want to recover
  start.time <- rnorm(n.rep, mean=16, sd=1)
  start.time[start.time < 10] <- 10
  start.time[start.time > survey.age - 5] <- max.age - 5
  stop.time <- rexp(n.rep, rate=alpha.Q) + start.time
  event <- (stop.time <= survey.age) + 0  # as numeric
  stop.time[stop.time > survey.age] <- survey.age
  sim.data <- rbind(sim.data,
                    data.frame(cohort=rep(cohort, n.rep),
                               wt=rep(0.5, n.rep),
                               start.time=start.time,
                               stop.time=stop.time,
                               event=event)
                   )
}
```

Chapter 6

Smoking cessation and recall error: estimation using a dynamic smoking model with census data

Abstract

Aim: To estimate smoking cessation and rates of recall error from former smokers to never smokers using census data. **Methods:** Data included: prevalence estimates from the New Zealand censuses in 1976, 1981 and 1996; cause-specific mortality rates; cause-specific mortality rate ratios from the literature; and retrospective smoking initiation rates from the 1996/97 New Zealand Health Survey. Prevalence was assumed to follow a multinomial distribution and a dynamic model was fitted using maximum likelihood for each single year birth cohort by sex. Similarly models were used to validate retrospective cessation rates from the 1996/97 New Zealand Health Survey. By assuming recall error was similar across ages, average cessation and initiation were estimated for younger respondents. **Results:** The patterns for rates of cessation between 1976–1981 and 1981–1996 were remarkably similar. Recall error by former smokers as never smokers was appreciable, being in the order of 1–3% per year. Retrospective cessation estimates were at variance with the census results. **Discussion:** The methods provide a robust mechanism for estimating different smoking parameters. Retrospective estimates may not be valid for more recent years.

6.1 Introduction

Two general approaches have been used recently to estimate smoking cessation rates. First, the hazard can be estimated using retrospective data and survival analysis methods. This approach was considered in Chapter 5), however there is an issue with whether retrospective estimates are valid.

An alternative approach is to use current status data. Mendez et al. (1998) fitted a dynamic smoking model to U.S. prevalence data to estimate the cessation rates. Although the model was arguably parsimonious, the data and estimated parameters were highly aggregated, limiting the validity or interpretation of any estimates (see Appendix B starting on page 279). As a consequence, the authors were unable to detect changes in rates. The authors assumed that smoking initiation after age 18 years was negligible, where good evidence suggests this may well not be true (see Chapter 5). The authors used rate ratios from an early report on the National Mortality Followback Study (Rogers and Powell-Griner, 1991), estimates from which have been substantially revised (Hummer et al., 1998). Irrespective of these reservations, however, the use of a dynamic model was advantageous, allowing both the explicit incorporation of differential mortality and straightforward projections of the process.

Chapter 3 describes a novel method using current status data to estimate rates of cessation from changes in prevalence combined exogenously with mortality differentials. This will be investigated further in Chapter 7 and Appendix B.

For a complete specification of the smoking model, there is a need to estimate the recall error by former smokers as never smokers (van de Mheen and Gunning-Schepers, 1994). If recall error is appreciable, then retrospective prevalence estimates of ever smoking will be biased. Moreover, some previous efforts to adjust for differential mortality will be biased (Burns et al., 1997b).

The aim of this chapter is to use data from the New Zealand Census of Population and Dwellings to establish a reference for the estimation of smoking cessation and recall error. This reference could be used to validate other estimation methods.

6.2 Methods

In outline, the methods review the smoking model from Chapter 3, describe the data sources, review the relevant theory, and describe the statistical methods. For a more complete theoretical development, see Chapter 3.

6.2.1 Review of the smoking model

A multi-state model was used to model smoking (see Figure 6.1) (Commenges, 1999; Hougaard, 1999). The model is described by a series of states, with initial values in each state, and a set of transition rates between the states. The transitions rates between the states were assumed to be piecewise homogeneous, being constant over one year periods, and Markov, with transitions being dependent only on sex, age and period.

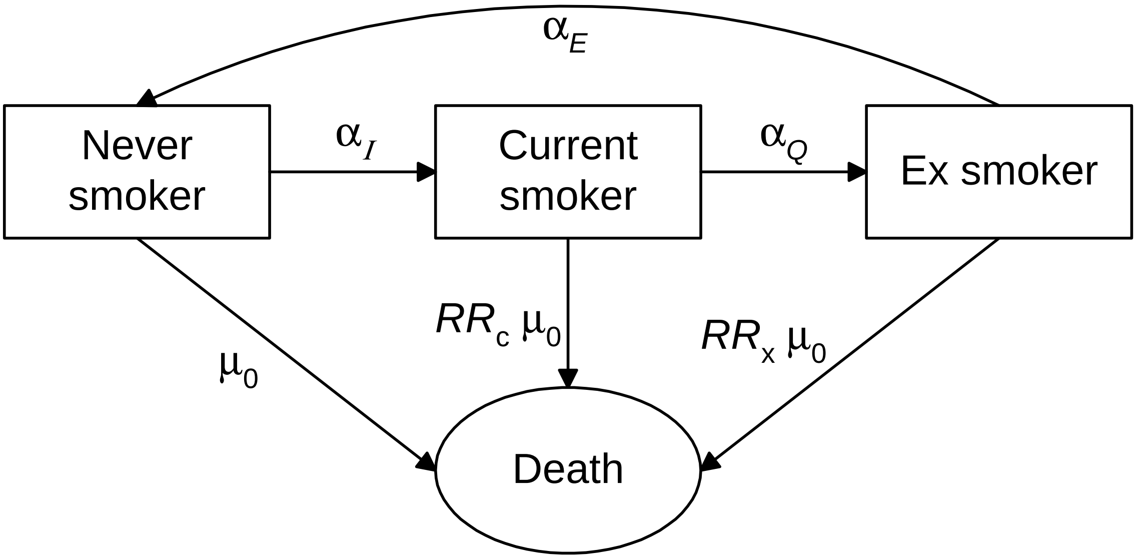


Figure 6.1: Graphical representation of a smoking model with uptake, cessation, recall error and differential mortality (Model 1)

States include never, current and former smokers, and death. There were transitions to death from each smoking state. There were also modelled transitions from never smokers to current smokers, and current smokers to former smokers, and from former to never smokers due to recall error. Reviewing the notation introduced in Chapter 3, let α_I, α_Q and α_E be the rates for initiation, cessation and recall error by former smokers as never smokers, respectively. Let μ_0, $RR_c\mu_0$ and $RR_x\mu_0$ be the mortality rates for never, current and former smokers, respectively. Migration rates were assumed to be independent of smoking status.

6.2.2 Data sources

From the New Zealand Censuses of Population and Dwellings for 1976, 1981 and 1996, respondents aged 15 years and over were asked whether they currently or had ever regularly smoked cigarettes. Regular smoking was defined as smoking one or more tobacco cigarettes per day.

Age-specific mortality rates were estimated from resident populations and deaths for 1976–1996 by five year age groups by sex. Data were provided by Statistics New Zealand. Age-specific mortality rates by single years of age were estimated by smoothing age-specific rates at the mid-points of the five-year age groups using thin plate smoothing splines (SAS Institute Inc., 1999).

To avoid assuming zero smoking initiation, smoking uptake was estimated from retrospective estimates from the 1996/97 New Zealand Health Survey. Initiation rates were available by single year of age by calendar year by sex. Details are given in Chapter 5.

Rate ratios for New Zealand were estimated using New Zealand condition-specific mortality rates together with condition-specific mortality rate ratios from the literature. Rate ratios were available by single year of age by calendar year by sex. Details are given in Chapter 4.

The rate ratios by one-year age groups were estimated by smoothing for age separately for current and former smokers, for males and females, for 1966 and 1986. The smoothing used thin plate splines with smoothing parameters estimated using generalised cross validation (`proc tpspline` in `SAS`). Then the age-specific rate ratios were linear interpolated between 1966 and 1986. In the absence of other information, it was assumed that the rate ratios did not change before 1966 or after 1986. Moreover, the rate ratios for those aged less than 35 years was assumed to be 1.00 for each group.

See Chapter 2 for additional information on the data sources.

6.2.3 Review of model theory

The basic approach is a generalisation of the (single decrement) cohort life table. Let the transition probability $P_{jk}(s,t)$ be defined as the probability of being in state k at time t conditional upon being in state j at time s. Moreover, let the transition rate $\alpha_{jk}(t)$ be defined as

$$\alpha_{jk}(t) = \lim_{\Delta t \downarrow 0} P_{jk}(t, t+\Delta t)/\Delta t \quad (j \neq k).$$

so that $\alpha_{jk}(t)\,\Delta t$ is the probability of a transition from state j to state k during the small interval $(t,\, t+\Delta t]$. Also let $\alpha_{jj} = -\sum_{k\neq j}\alpha_{jk}$. For a single decrement life table, where j is the live state and k is the death state, $P_{jk}(s,s+1)$ would be the one-year death probability q, $P_{jj}(s,t)$ would be the probability of survival from time s to time t, and $\alpha_{jk}(t)$ would be the mortality rate m.

Using matrix notation, let $\boldsymbol{P}$ and $\boldsymbol{\alpha}$ denote a matrix of transition probabilities and transition rates, respectively. Assuming the transitions are piecewise homogeneous Markov processes, then Chapman-Kolmogorov's equation and Kolmogorov's equation (Chiang, 1980; Andersen et al., 1993) give

$$\begin{aligned}\boldsymbol{P}(s,t) &= \prod_{u=s}^{t-1} \boldsymbol{P}(u,u+1) \\ &= \prod_{u=s}^{t-1} \exp\left(\boldsymbol{\alpha}(u)\right).\end{aligned}$$

As a summary of this relationship, given the transition rates by single year, we can mechanically calculate the transition probabilities at a later time. The matrix exponentiation requires numerical solution using standard routines for ordinary differential equations.

For the model, $\boldsymbol{\alpha}$ for the live states is represented by

$$\boldsymbol{\alpha} = \begin{bmatrix} -\mu_0 - \alpha_I & \alpha_I & 0 \\ 0 & -RR_c\mu_0 - \alpha_Q & \alpha_Q \\ \alpha_E & 0 & -RR_x\mu_0 - \alpha_E \end{bmatrix}.$$

The death state has not been explicitly included, because its exclusion simplifies the model and does not affect the model estimates.

6.2.4 Estimation

The prevalence for current and former smokers is assumed to follow a multinomial distribution. The parameters to be estimated are initial prevalence for current and former smokers, the rates of cessation and recall error. The regression equation and likelihood are given in Section 3.6.4.

For a given sex and birth cohort c:

1. Select initial values for the parameter values

2. Maximise the log likelihood over the dynamic model, where

    (a) Initialise the dynamic model

        i. Set time t as the end of the intercensal period

        ii. Set age a $(= t - c)$

        iii. Calculate prevalence at time t from the parameters

(b) Step through the dynamic model, moving from the end to the beginning of the inter-censal period

   i. Extract total mortality rate, rate ratios for current and former smokers, and recalled initiation and cessation rates

   ii. Calculate the initiation, cessation and recall error rates from the parameters [varies dependent upon model]

   iii. Calculate never smoker mortality rate

   iv. Calculate survival from the (cumulative) transition probability matrix

   v. Calculate initiation and cessation rates adjusted for survival

   vi. Calculate the transition intensity matrix

   vii. Calculate the (step) transition probability matrix

   viii. Update the state vector and the cumulative transition probability matrix, and calculate prevalence

   ix. Update year $(= t - 1)$ and age $(= a - 1)$

(c) Return the model log likelihood.

The negative log-likelihood was minimised using `nlm()` in `R`, which is a non-linear minimisation routine that employs a Newton-type algorithm. The start-point for the minimisation was found using a grid search. The covariance matrix was estimated from the inverse of the Hessian matrix (Venables and Ripley, 1999).

Given all cause mortality rate ratios by smoking status together with mortality rates, and initial conditions for each smoking state, the initial model was fitted to estimate the quit rate for each five-year cohort and sex combination. This was done for the interval between the 1976 and 1981 Censuses, and for the interval between the 1981 and 1996 Censuses. The model fitting involved finding the value of the quit rate that gave the correct current smoking prevalence at the Census at the end of the interval.

Models of interest included:

Model 1A For ages 25 years and over, initiation rates were taken from retrospective data, with adjustment for differential survival, and average cessation and recall error rates were estimated.

Model 1B As per Model 1a, except that the cessation rates were taken from retrospective data and a scale factor for the cessation rates was estimated.

Model 1C For ages under 25 years, recall error rates were assumed known (0.02/year for females and 0.01/year for males) and average initiation and cessation rates were estimated. This was done for the 1976–1981 intercensal period.

The never smoker mortality rate was estimated using

$$\mu_0 = \frac{\mu}{\pi_c(RR_c - 1) + \pi_x(RR_x - 1) + 1}.$$

Model 1A follows Model 1 described in Chapter 3, with initiation rates estimated from retrospective estimates from the 1996/97 New Zealand Health Survey and the parameters for cessation and former smoker recall error being estimated from the data. The estimated hazard rates were averages for the intercensal periods.

Model 1B is similar, except that the baseline for cessation was taken from the 1996/97 New Zealand Health Survey and a scale factor was estimated, where estimated hazard for cessation equals baseline hazard times the scale factor. Again, the scale factor was a constant across the intercensal period.

Model 1C assumes recall error for younger respondents is similar to the older respondents. The main interest is characterising the functional form for smoking cessation for the younger respondents. Estimation for youth is difficult because of the rapid changes in behaviour which will be measured differentially dependent upon the study design, affecting fitting Model 1A and Model 1B. A sensitivity analysis was carried out for Model 1C, re-fitting all models for a 1% increase in the recall error rate.

6.3 Results

6.3.1 Change of prevalence

Over the period 1976–1996, there has been a considerable decline in the proportion of adults smoking tobacco (Figure 6.2). For adults aged 15 years and over, in 1976 39.6% of males and 31.7% of females regularly smoked cigarettes, compared 24.8% of males and 22.8% of females in 1996. Care should be taken in the interpretation of these changes as the two intercensal periods were of different lengths.

This reduction has not been consistent between sexes or between ages. In absolute and relative terms, males have changed their behaviour more quickly. As an example of differential changes by age groups, there was limited change for smoking prevalence for younger and older females between 1976 and 1981, while females aged 25–54 years had more marked declines.

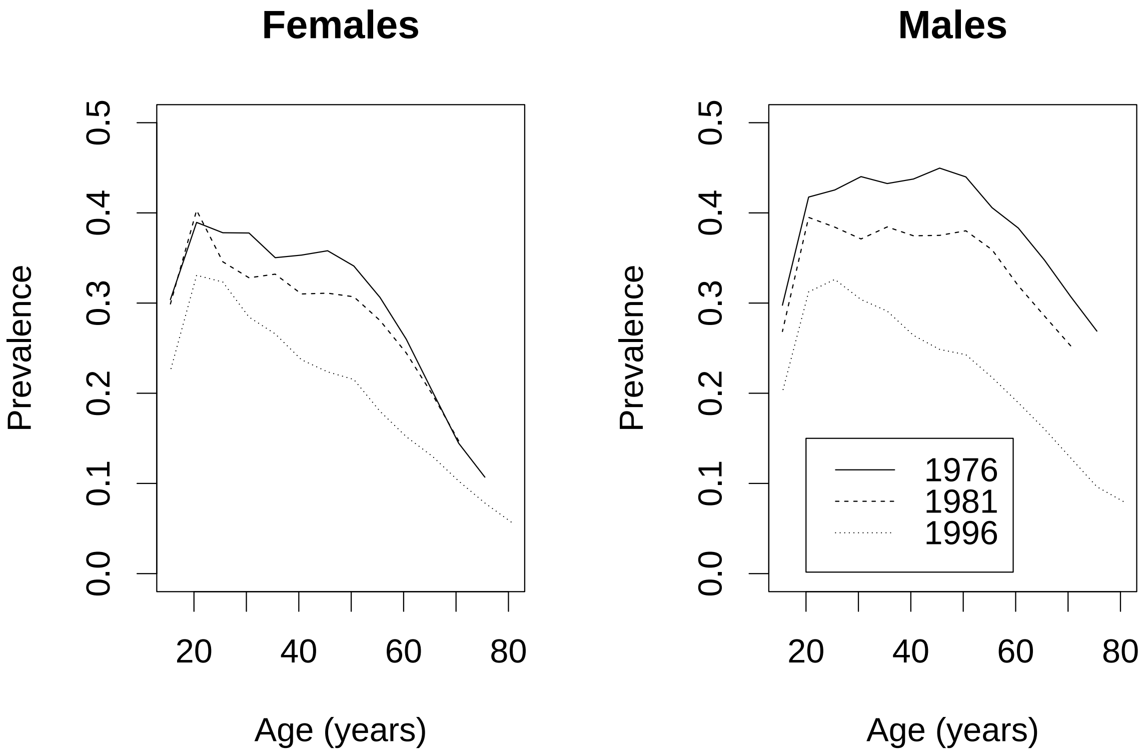


Figure 6.2: Current smoking prevalence among New Zealand adults aged 15 years and over, New Zealand Census of Population and Dwellings (1976, 1981 1996)

6.3.2 Model 1A: Average cessation and recall error

As an example of the model fitting, consider females born in 1930–1931. In 1976 the women were aged 46–47 years and the prevalence of current and former smoking was 35.8% and 12.4%, respectively. Five years later, the prevalence of current and former smoking was 31.1% and 15.3%, respectively, so that ever smoking decreased by 1.3% across the intercensal period. By fitting the model, the estimated cessation rate was 0.029 per year (95% confidence interval: 0.025, 0.033) and the estimated recall error rate was 0.028 per year (95% confidence interval: 0.017, 0.039). For the transition probability matrix $\boldsymbol{P}$ for the intercensal period 1976–1981, see Table 6.1.

Note that approximately 12% of former smokers in 1976 were expected to change their recall to being never smokers by 1981. Survival fractions through the period, calculated by summing across the rows from the transition matrix, for those that were never, current and former smokers at the beginning of the period were 0.985, 0.975 and 0.981, respectively. The transition probability from former to current smokers ($P_{xc}(1976, 1981) = 0.00001$) is a negligible artefact representing uptake by former smokers who have changed their recall to never smokers. Otherwise, the

	N	C	X
N	0.985121	0.0001901	0.0000161
C	0.007832	0.8454823	0.1212435
X	0.118187	0.0000100	0.8630664

Table 6.1: Transition probability matrix for New Zealand women born 1930–1931 for the intercensal period 1976–1981

transition probabilities have useful interpretations.

Similar calculations were performed for each sex and birth cohort. Estimates for cessation rate are shown in Figure 6.3. There was a surprisingly close agreement in average cessation rates between males and females and between the two intercensal periods. The cessation rates were lowest around 40 years of age, with a linear rise with increasing age. For younger ages, there was a suggestion that rates may also have been high.

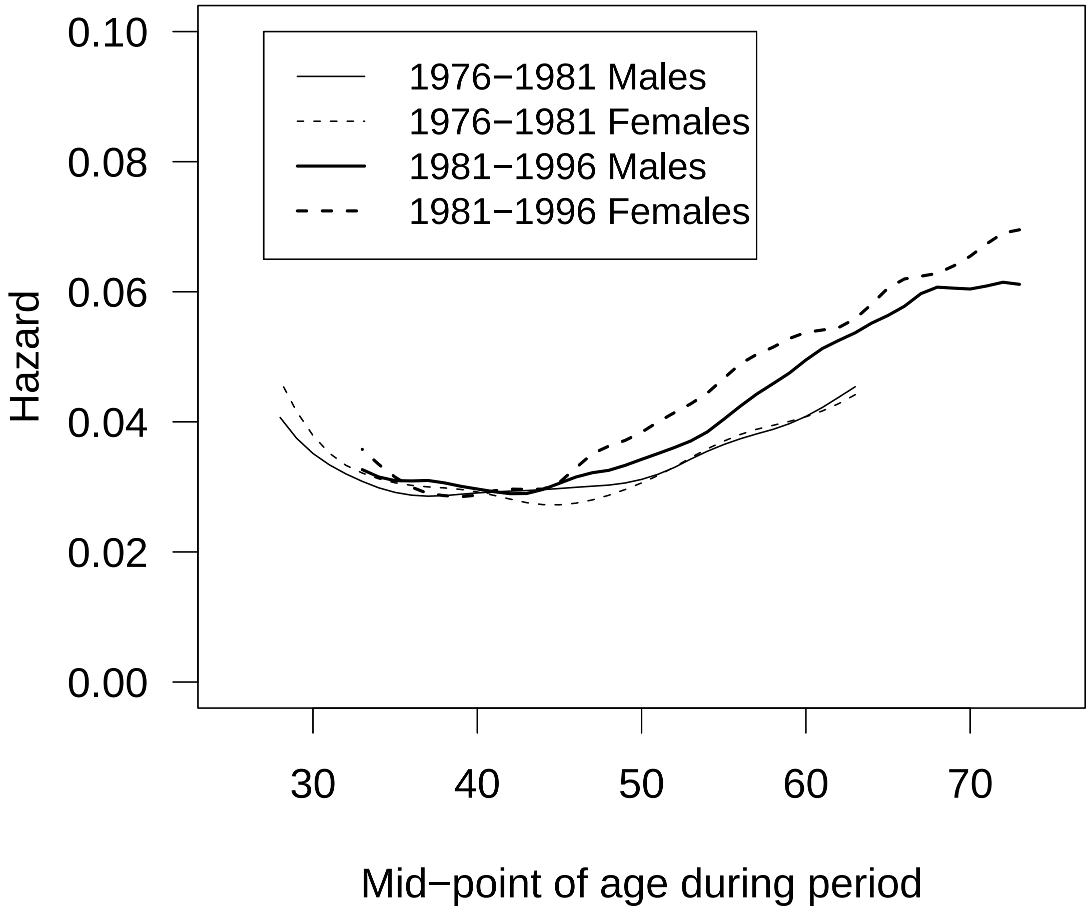


Figure 6.3: Average intercensal hazard rates for smoking cessation, Model 1A

The average rates of recall error by former smokers were higher for females than males (Figure 6.4). The standard errors for the recall error rate estimates were larger than those for smoking cessation (results not shown), which may partially explain the more variable rates for females. Recall error by females for 1976–1981 appears to be higher than rates for 1981–1996. However males did not exhibit the same shift between the two intercensal periods.

The "kicks" in the tails for the younger respondents were also observed for the cessation rates. This may be an indication of rapid changes or of model instability. For older ages, the rates tended to be relatively stable by age. One possible exception is a slight drop in recall error for males aged 50 years and over during 1976–1981.

Average level for males was approximately 0.01 per year, while for females it varied between 0.02 and 0.03 per year.

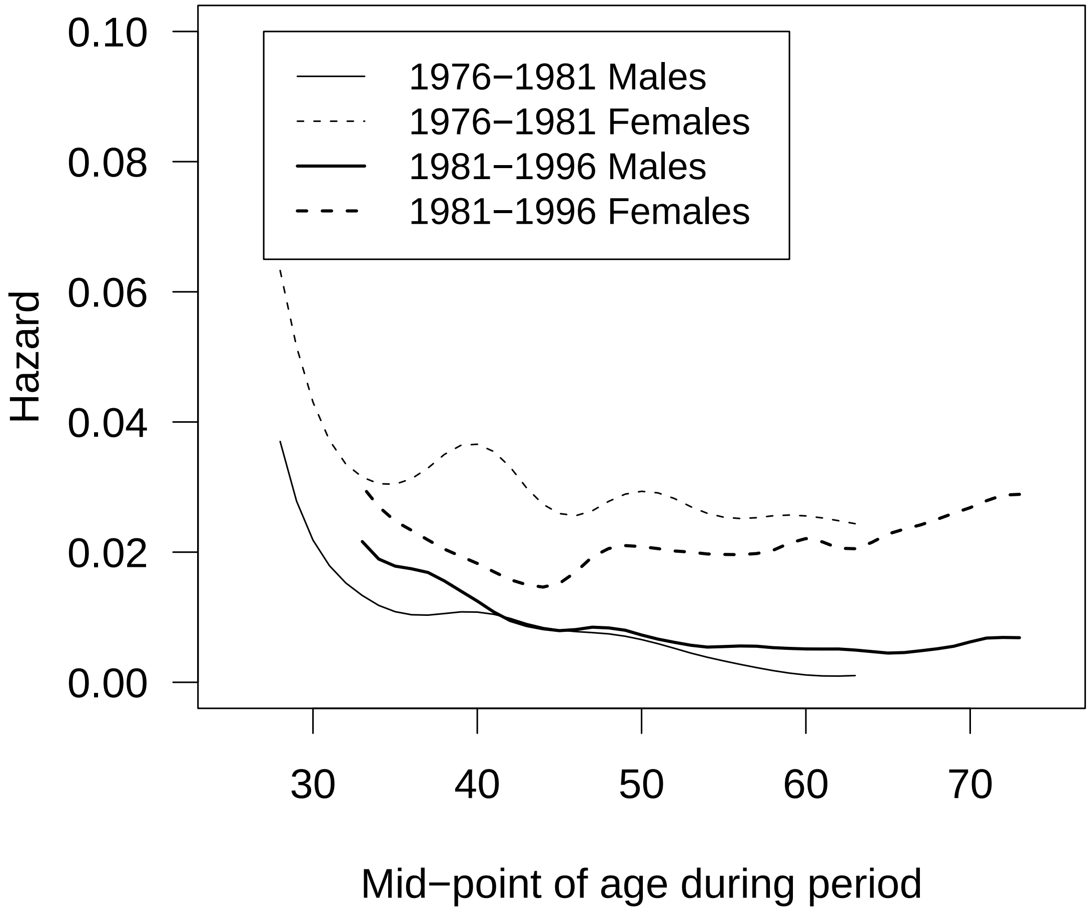


Figure 6.4: Average intercensal hazard rates for former smoker recall error, Model 1A

6.3.3 Model 1B: Validation of retrospective cessation estimates

The same data were re-fitted for a model incorporating retrospective estimates for cessation rates together with a multiplicative scale factor for cessation. The scale factors for cessation rates from the 1996/97 New Zealand Health Survey are shown in Figure 6.5. The results suggest that the level for the retrospective estimates are similar to or need to be scaled upwards from the Census results.

The age-specific scale factors were similar between the two intercensal periods for males and females, which may be explained by the smoothing used to estimate the retrospective cessation rates.

However the age-specific patterns varied considerably between males and females. Retrospective estimates for males aged 40–49 years had little bias, while estimates for females of a similar age needed to be scaled upwards by 25–50%. Moreover, estimates for older and younger males required inflation, whereas estimates for females at the oldest and youngest ages tended to have no bias.

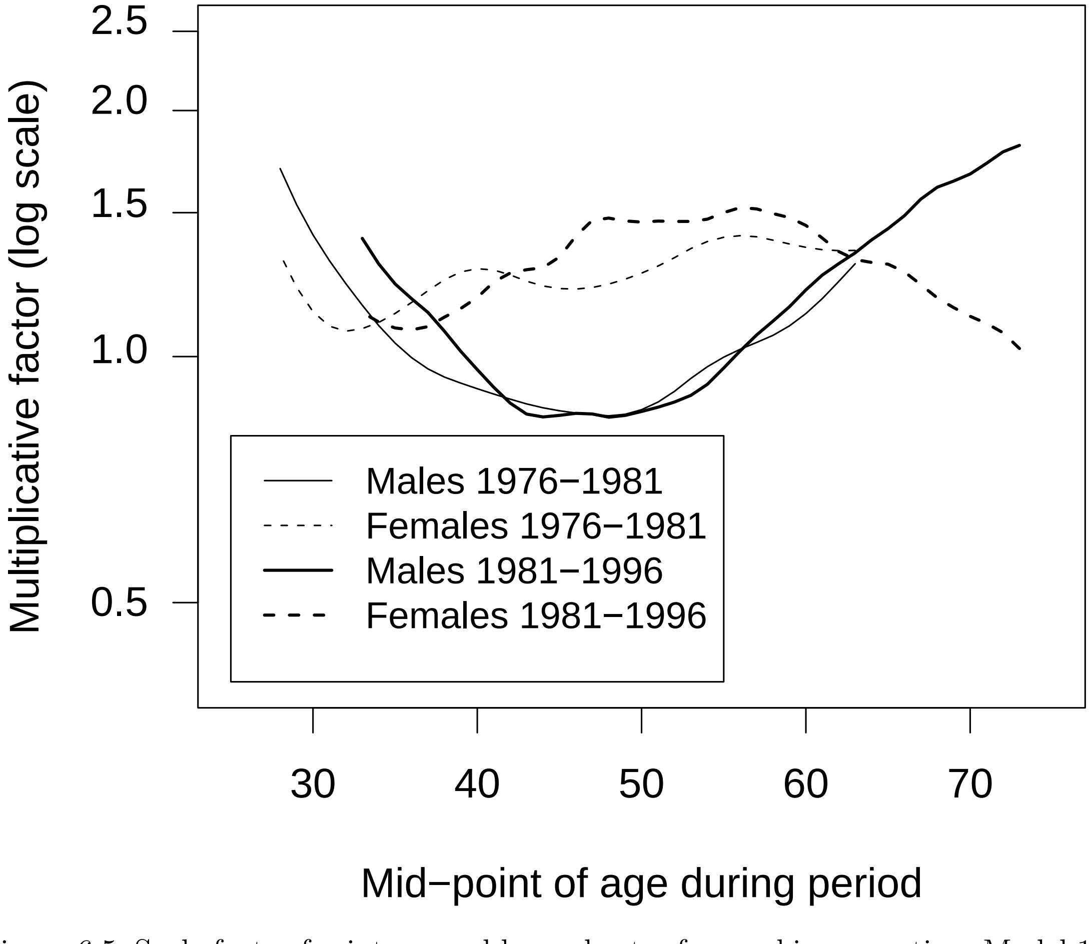


Figure 6.5: Scale factor for intercensal hazard rates for smoking cessation, Model 1B

The estimates for recall error from Model 1A and Model 1B were similar, so the latter results are not reported here.

6.3.4 Model 1C: Youth cessation estimates

At younger ages, the fits for Model 1A and Model 1B were not stable and were increasingly sensitive to the choice of retrospective initiation rates. One solution for this problem was to fix the recall error rates based on the older age groups.

For the resultant model, the estimated average cessation rates did not vary considerably by age (Figure 6.6). Rates were lowest in the late teens, peaked at age 21 years and were stable thereafter. Note that the results were averages, which would tend to smooth short-term peaks. Male rates were slightly lower than the female rates, however the functional forms were similar. The initiation rates in Figure 6.6 were not used for later analysis because they were averages for the intercensal period.

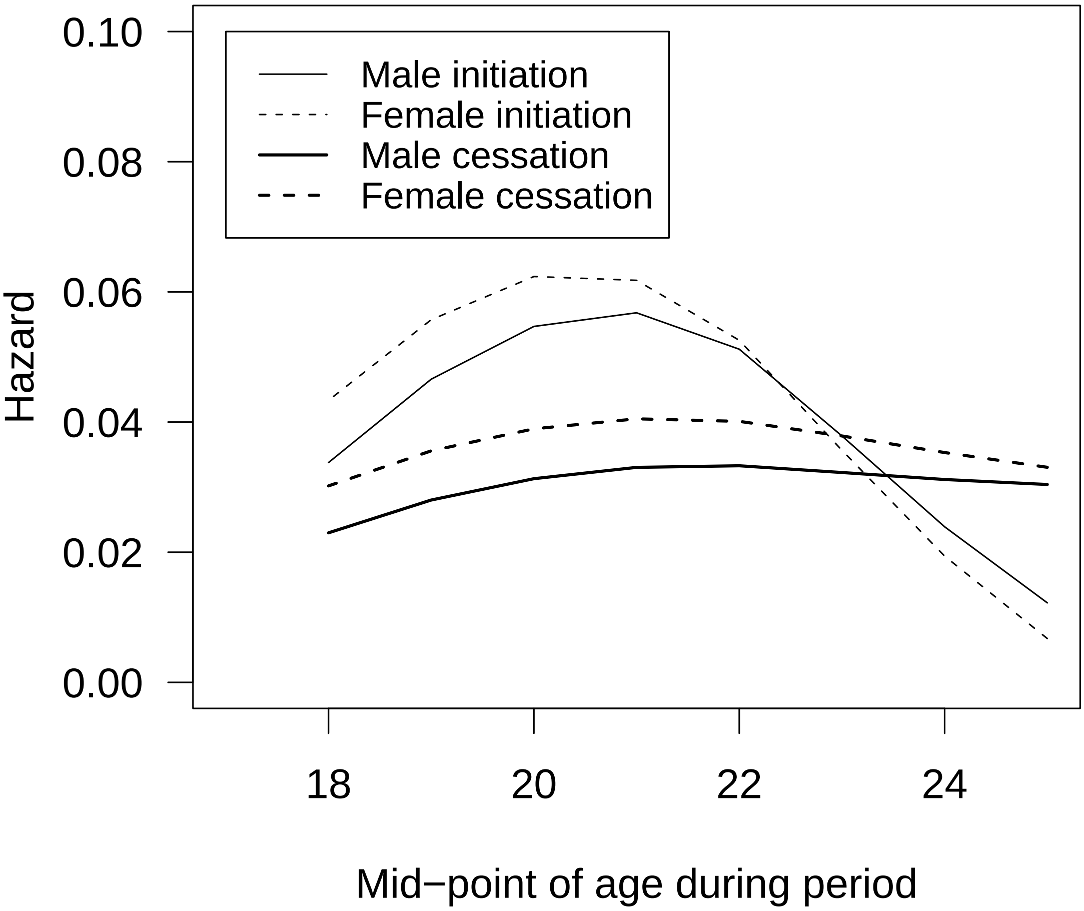


Figure 6.6: Average hazard rates for smoking initiation and cessation among youth between 1976 and 1981 New Zealand Census of Population and Dwellings

For a sensitivity analysis, the recall error rates were increased by 1% and the model was re-fitted. The impact was a small proportional increase in the cessation rates for males and for females by 0.08% and 0.15%, respectively.

6.4 Discussion

In summary, smoking parameters were carefully estimated from a multi-state model fitted using census data supplemented by retrospective initiation rates and mortality rates. The functional form for cessation rates was stable over the two intercensal periods. Validation of retrospective cessation rate estimates suggests that the estimates may be biased. An important finding is that recall error by former smokers as never smokers was 1–2% per annum.

6.4.1 Limitations

There were several limitations to the analysis.

First, parameter estimates for Model 1A and Model 1B were imprecise at younger ages because of rapid changes in behaviour. One result was that the likelihood surface became flat and difficult to fit. The proposed solution was to restrict Models 1A and 1B to ages 30 years and over and model the younger ages by fixing the recall error rate and estimating the initiation and cessation rates (Model 1C).

Second, recall of smoking behaviour from a population census has certain limitations, including differential systematic under-reporting over time and differential non-response rates. Jackson and Beaglehole (1985) used data from the 1976 and 1981 censuses to show that under-reporting was consistent over time. There was evidence for increasing item non-response between the 1981 and 1996 Census, which did not affect these results after post-stratification, but which may be of concern in later censuses.

Third, estimates may potentially be affected by differential migration by smoking status. This could be assessed by a sensitivity analysis using differential migration rates. This would require a full demographic model for population, using a cohort component model with forward-reverse survival methods (Shyrock et al., 1976).

Fourth, direct maximisation of the log-likelihood had poor convergence properties (Seber and Wild, 1989). Alternative approaches include assuming approximate normality for prevalence and using non-linear least squares or, preferably, using iteratively re-weighted least squares. Direct maximisation had the benefit of computational simplicity, and convergence was handled by use of a grid search.

6.4.2 Implications

The evidence that males and females required different adjustment of the retrospective estimates is cause for concern. One possible explanation is that the retrospective estimates are sensitive to the choice of fitting parameters and may not be reliable.

The main strength of this analysis is the use of Census data to model smoking behaviour. Few assumptions have been made, suggesting that this approach may be near optimal for estimating transition probabilities for such changes in smoking behaviour. A large cohort study with long follow-up would potentially provide better data on recall error, however such a study would be limited by cost and precision.

Two possible extensions to the approach would be to validate these methods in another population, and to use a full Bayesian model for variance estimation.

6.A Census results

Results from the three New Zealand censuses by census, sex and cohort are shown in Table 6.2.

		1976			1981			1996		
Sex	Birth cohort	Pop	π_c	π_x	Pop	π_c	π_x	Pop	π_c	π_x
Female	1906–1910	56643	0.2020	0.1460	48369	0.1472	0.1666	–	–	–
	1911–1915	66304	0.2599	0.1473	60171	0.1985	0.1803	34872	0.0573	0.1967
	1916–1920	70639	0.3062	0.1460	66645	0.2447	0.1809	48729	0.0783	0.2273
	1921–1925	77535	0.3411	0.1366	73272	0.2806	0.1742	61899	0.1027	0.2537
	1926–1930	78717	0.3580	0.1271	75036	0.3070	0.1571	67785	0.1300	0.2533
	1931–1935	74732	0.3532	0.1207	72246	0.3110	0.1486	67845	0.1519	0.2425
	1936–1940	87380	0.3503	0.1188	83925	0.3101	0.1432	79821	0.1799	0.2229
	1941–1945	98651	0.3777	0.1285	95517	0.3321	0.1607	93363	0.2153	0.2371
	1946–1950	123472	0.3780	0.1250	118779	0.3280	0.1612	120936	0.2237	0.2252
	1951–1955	127931	0.3894	0.1034	119940	0.3458	0.1504	129600	0.2372	0.2150
	1956–1960	–	–	–	145920	0.4029	0.1189	145920	0.2658	0.2206
Male	1906–1910	49805	0.3476	0.3971	38463	0.2526	0.4640	–	–	–
	1911–1915	60924	0.3832	0.3612	51876	0.2857	0.4403	20409	0.0801	0.5055
	1916–1920	66547	0.4057	0.3454	59703	0.3193	0.4192	33561	0.0963	0.5344
	1921–1925	78166	0.4399	0.3161	73164	0.3591	0.3849	51759	0.1282	0.5390
	1926–1930	84161	0.4496	0.2679	79104	0.3801	0.3238	65184	0.1603	0.4805
	1931–1935	77899	0.4376	0.2320	75054	0.3751	0.2826	67419	0.1895	0.4170
	1936–1940	89717	0.4325	0.2018	85026	0.3746	0.2485	78789	0.2175	0.3622
	1941–1945	100847	0.4402	0.1793	95931	0.3844	0.2284	93348	0.2427	0.3334
	1946–1950	125668	0.4254	0.1477	118779	0.3711	0.1989	120249	0.2486	0.2899
	1951–1955	131644	0.4175	0.1046	118515	0.3842	0.1582	125436	0.2642	0.2530
	1956–1960	–	–	–	139293	0.3950	0.1062	139293	0.2911	0.2133

Table 6.2: Census smoking data, New Zealand, 1976, 1981 and 1996

Chapter 7

Cessation rate estimation for Australia using current status prevalence data

Abstract

Aim: To estimate the rate of cessation using the prevalence of current status prevalence data. **Methods:** The cessation rate was calculated from the rate of decline of current smoking prevalence less the mortality differential between current smokers and the total population plus an adjustment for smoking uptake. Smoking prevalence was modelled using non-parametric local likelihood methods and parametric generalised linear models. **Results:** There was reasonable agreement between methods based on local likelihood estimation and generalised linear models for older ages. At younger ages, where changes in behaviour were more rapid, estimates were less consistent between the two methods. There was an increase in cessation rates by age and with years at the end of the period 1974–1995. **Discussion:** A simple method was used to estimate the rate of cessation using current status prevalence data that avoided more complicated modelling.

7.1 Introduction

From the analysis of Census data in Chapter 6, there are concerns that retrospective estimates of cessation rates may be biased. This concern has previously been discussed by Diamond and McDonald (1991), who urged caution in the analysis of retrospective data and suggested more wide-spread use of current status data.

Mendez et al. (1998) recently suggested using current status smoking data to estimate smoking cessation rates. The authors used a dynamic smoking model. There are concerns with their approach, as their results are sensitive to model assumptions, their data poor, and their conclusions are at odds with observed data. See Appendix B.

An extension to the approach by Mendez et al. (1998) would be to include states for never, current and former smokers, and to include initiation rates and differential mortality. This is similar to the approach in Chapter 6, which also estimates for the rate of recall error and uses a multinomial distribution.

A simpler approach is to use a regression approach on current status smoking prevalence data supplemented with mortality and initiation rates (see Section 3.6.3 for the relevant development). This is the approach used in this chapter.

7.2 Methods

7.2.1 Data sources

Australian data sources include the prevalence of current and former smokers provided by the Anti-Cancer Council of Victoria (ACCV). Age-specific mortality rates were available from vitals. For further detail, see Chapter 2. Smoking initiation rates were taken from retrospective estimates (see Chapter 5). All cause mortality rate ratios were available from Chapter 4.

7.2.2 Review of theory

A theoretical development for estimation of the cessation rate from current status data is given in Section 3.6.3. The main relationship is given in Equation (3.22) on page 80, repeated here:

$$\alpha_Q = -\frac{\mathrm{d}\pi_c}{\mathrm{d}t}/\pi_c - (RR_c\mu_0 - \mu) + \alpha_I\frac{\pi_n}{\pi_c}$$

where t is time, α_Q and α_I are the rates of smoking cessation and initiation, respectively, π_c and π_x are the prevalence of current and former smokers, respectively, RR_c is the all cause mortality rate ratio for current smokers, and μ and μ_0 are the mortality rates for the total population and never smokers, respectively. This relationship states that the cessation rate for a birth cohort is equal to the rate of decline of current smoking prevalence less some mortality differential, plus an adjustment for smoking uptake.

By further defining RR_x as being the all cause rate ratio for former smokers and using Equation (3.1), the mortality differential can be calculated from

$$RR_c\mu_0 - \mu = \mu\left[\frac{RR_c}{\pi_c(RR_c - 1) + \pi_x(RR_x - 1) + 1} - 1\right].$$

At older ages, initiation rates are negligible and can be ignored. However, for this presentation, retrospective estimates of the rates were included to give some indication of the cessation rates at younger ages.

7.2.3 Estimation of the rate of decline of prevalence

The remaining analytical challenge was to estimate the rate of decline of current smoking prevalence. Two approaches were used. First, local likelihood was used to provide non-parametric estimates of the rate of decline. As noted in Section 3.6.3, the implementation of local likelihood estimation (`locfit`) requires a logit link, so that Equation (3.27) will be required. The partial derivatives for the linear predictor were provided as output from `locfit` using the `deriv=` statement.

Second, generalised linear models were used to provide parametric estimates (McCullagh and Nelder, 1989), with the functional form as suggested from the non-parametric approach. For this second approach, a log link can be used, so that point and variance estimates are described by Equations (3.25) and (3.26), respectively. Models were selected based on Akaike's Information Criterion. The partial differentials were estimated numerically.

Both approaches assumed that the observed proportions had a binomial distribution. The derivative for a cohort was calculated using the partial differentials, as suggested by Equation (3.9).

In order to obtain estimates to age 75 years, open age groups were used, using the tuple of the average age in the open age group and the prevalence for analysis. A validation of this approach using 1996 New Zealand Census data by single year of age is given in Appendix 7.A on page 179.

7.2.4 Calculation of the rates of cessation

Rates of decline of current smoking prevalence were estimated by single year of age and single year of calendar period. Prevalence for current and former smokers were similarly estimated using local likelihood estimation. Finally, these estimates were combined with vitals and rate ratios to estimate the rate of cessation.

7.3 Results

Smoothed prevalence surfaces of current smoking were shown as perspective plots in Figure 1.10 on page 13 and Figure 1.11 on page 13. The rate of change of prevalence can be interpreted as the slope (on the diagonal) on these surfaces expressed as a proportion of the prevalence.

For generalised linear models, the best model for male current smoking prevalence included a cubic polynomial for age, a linear effect for year, and an interaction between age and year. In R modelling syntax, the model was fitted using

```
glm(cbind(current, total - current) ~ poly(age, 3) +
            year + age : year,
            data = males, family = binomial(link = "log"))
```

where current and total were the weights for current smokers and total population for a cell, respectively, and year and age are covariates for year and age, respectively.

The model for females included a cubic for age and quadratic for year, with interactions between linear age and year, linear age and the square of year, and the square of age and linear year. Again expressed in R modelling syntax, the model was fitted using

```
glm(cbind(current, total - current) ~ poly(age, 3) +
            poly(year, 2) + age : poly(year, 2) + I(age ^ 2) : year,
            data = females, family = binomial(link = "log"))
```

The inclusion of the interaction of quadratic age and linear year was equivocal, but the final estimates did not change appreciably with either model.

Estimates of the rates of decline estimated from the local likelihood models using locfit are compared with estimates from the generalised linear models in Figures 7.1 and 7.2. For females, the pattern at older ages was similar, with increasing cessation at older ages, with rates increasing for more recent years. For males, the rates at older ages from the generalised linear model rose more slowly than the local likelihood estimates. The pattern at younger ages differs between the two models, which may be explained by the models trying to fit for more rapid changes in smoking behaviour. For males, another explanation may be that the simple functional form used in the generalised linear model may poorly represent any rapid changes.

The local likelihood estimates for the rates of decline of current prevalence were used in later analysis due to their flexible functional form. However any formal analysis would require the use of the generalised linear models.

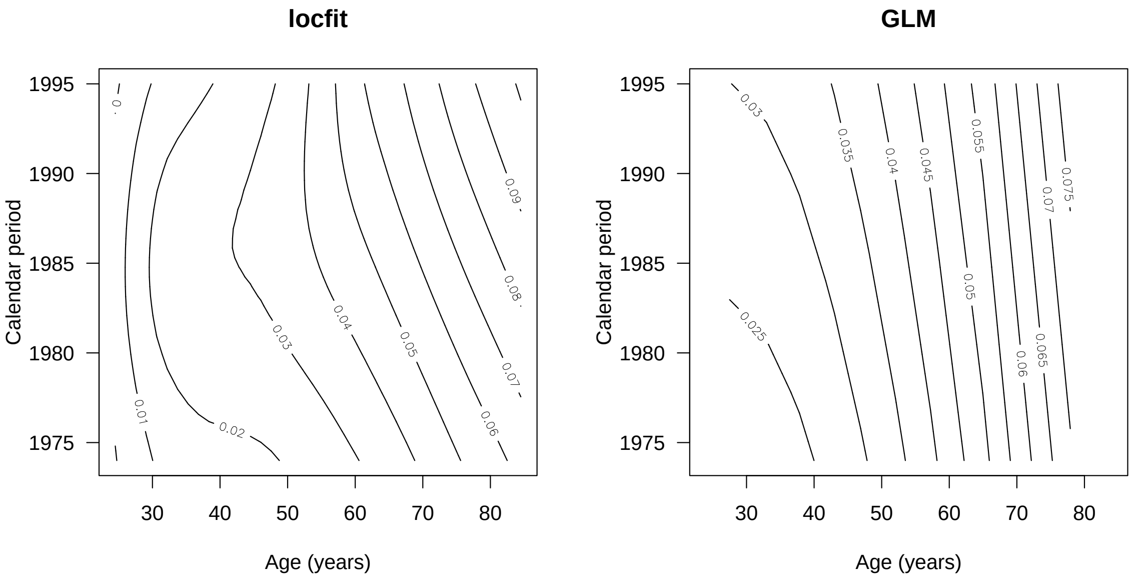


Figure 7.1: Estimated rate of decline of current smoking prevalence by age and period for local likelihood and generalised linear models, Australian males

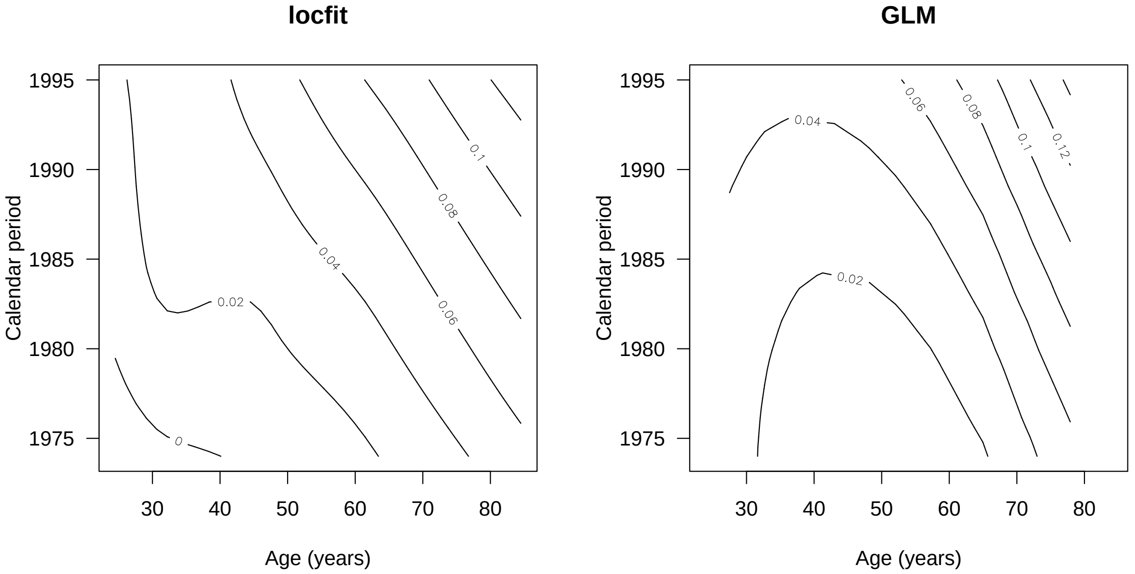


Figure 7.2: Estimated rate of decline of current smoking prevalence by age and period for local likelihood and generalised linear models, Australian females

Estimates of the rate of cessation based on the local likelihood estimation are shown in Figures 7.3 and 7.4. In comparison with the rates of decline of prevalence, the rates of cessation have increased at younger ages, which can be attributed to adjusting for the initiation rates. The effect due to initiation is larger for females than for males. The cessation rates are lower than the rates of decline of prevalence at older ages due to differential mortality. This effect is limited before 60 years of age.

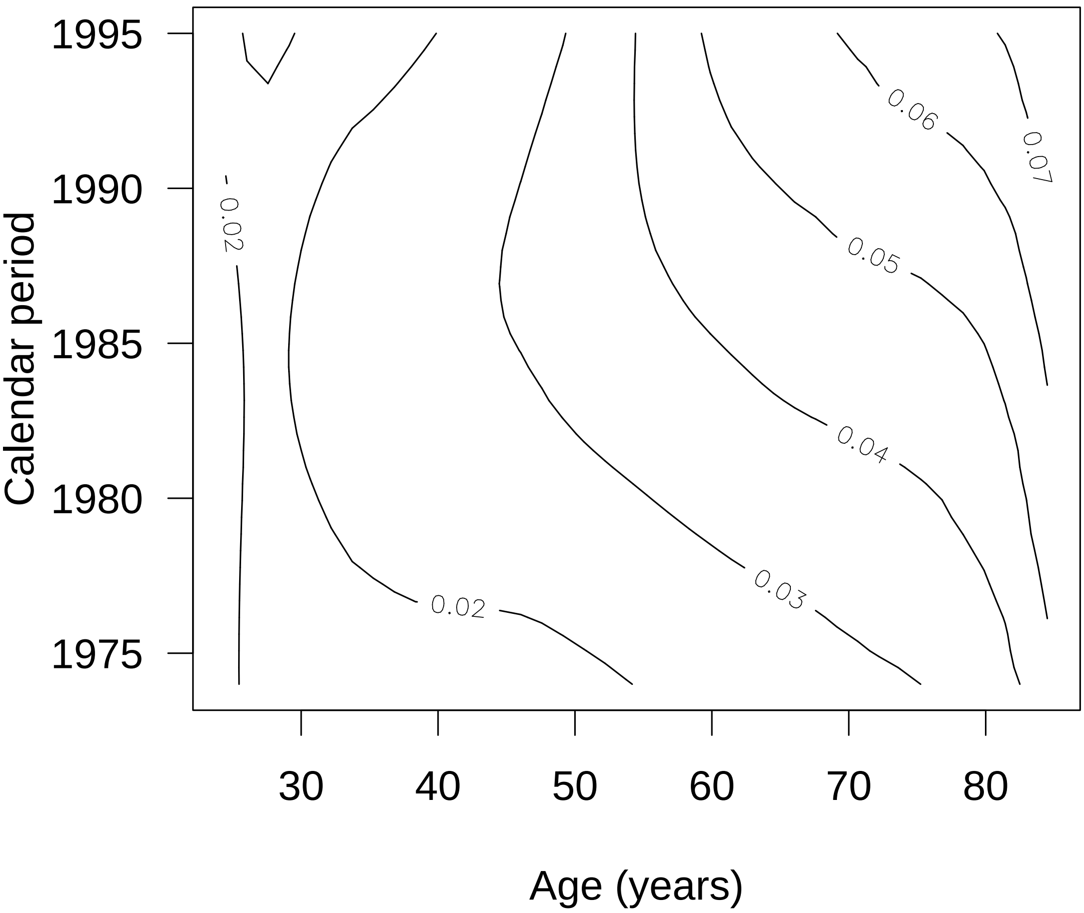


Figure 7.3: Estimated smoking cessation rates by age and period, Australian males 1974–1995

7.4 Discussion

In summary, current status smoking prevalence data together with supplementary data have been combined to estimate the rate of smoking cessation. There was reasonable agreement between methods based on local likelihood estimation and generalised linear models for older ages. At younger ages, where changes in behaviour were more rapid, estimates were less consistent between the two methods. There was a distinct increase in cessation rates by age and with years at the end of the period 1974–1995.

There are several limitations to this method. First, estimation of the rate of decline of smoking prevalence is difficult. Estimation of prevalence is more precise than estimating the rate that prevalence changes: using a geometric argument, it

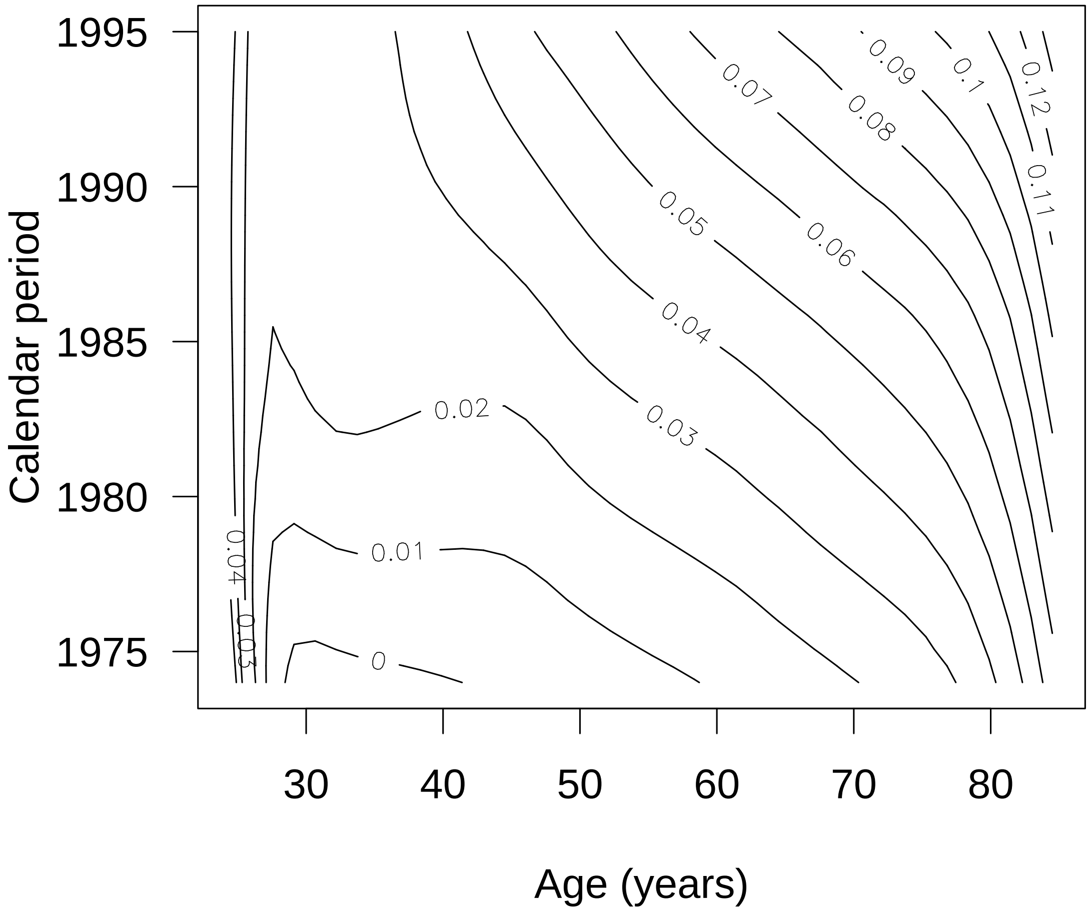


Figure 7.4: Estimated smoking cessation rates by age and period, Australian females 1974–1995

is easier to measure the value of a function at a point than it is to measure the slope at that point. As a counter-point, differences in the sampling frames mean that the *validity* of any comparison of prevalence between studies may be less than any comparison of rates of change between studies.

Moreover, from a technical perspective, derivative estimation is dependent on the bandwidth (for local likelihood) and on the degrees of freedom (see Loader, 1999). Taking extreme cases, a constant function with one degree of freedom will have no slope, while a function that interpolates between points would have a sharp and varying slope.

Second, variance estimation for Equation (3.22) is involved, particularly as the prevalence data are included in ratios and in sums. One possible approach is to use parametric forms using generalised linear models together with the delta method (see Appendix B). A dynamic modelling approach would allow inclusion of the variation due to the prevalence estimates, but no other sources of variation would be included.

Current status estimates of cessation have the advantage over retrospective estimates of being based on prevalence estimates, hence the smoking model estimates are expected to be more consistent. It is not possible to compare the cessation estimates from this chapter with the census estimates from Chapter 6, as they come from different populations.

Cessation rate estimation for New Zealand using current status data is difficult due to data constraints. In particular, the available time series for current smoking prevalence are highly aggregated by age and the prevalence proportions of former smokers are not available. Valid estimates for New Zealand would therefore need to rely on retrospective data coupled with validation using census data. See Chapter 6 for further discussion.

The approach followed here is considerably simpler than a dynamic modelling approach such as used by Mendez et al. (1998). Further simplifications could be to ignore differential mortality and uptake by not considering the oldest and youngest age groups. This method provides a simple approach that may deserve broader application.

7.A Modelling for open age groups

In order to be able to model smoking prevalence when some of the data are available by open age groups, the following observations were made. Let prevalence of current smoking at a given point in time be $\pi_c(a)$ for age a. Moreover, let the age-specific population distribution be represented by a probability density function $w(a)$. Then the average prevalence for an age interval $[a_0, a_1)$ is

$$\bar{\pi}_c = \left[\int_{a_0}^{a_1} \pi_c(a)w(a)\mathrm{d}a\right] / \left[\int_{a_0}^{a_1} w(a)\mathrm{d}a\right].$$

and the average age in the interval is

$$\bar{a} = \left[\int_{a_0}^{a_1} aw(a)\mathrm{d}a\right] / \left[\int_{a_0}^{a_1} w(a)\mathrm{d}a\right].$$

If $\pi_c(a)$ is approximately linear in the interval then

$$\bar{\pi}_c \approx \pi_c(\bar{a}).$$

This approximation was found to be reasonable under mild non-linearity such as would be seen at older ages. Prevalence for an open age group can then be represented by $(\bar{a}, \bar{\pi}_c(a))$, using the prevalence for the age group as already available

	Age group (years)			
	70+	75+	80+	85+
Males	77.1	80.8	84.5	88.5
Females	78.4	81.8	85.3	89.1
Total	77.9	81.4	85.0	88.9

Table 7.1: Mean ages for open age groups, 1996 New Zealand Census

and the mean age for the open age group. This approach is a generalisation of the approach that uses mid-points to model for closed age groups, which assumes that the mean age is at the middle of the interval.

Data were available from the 1996 New Zealand Census of Population and Dwellings by single year of age and smoking status. The average ages in the population, irrespective of smoking status, for different open age groups are shown in Table 7.1.

One form of validation is to see whether the mean value is obtained at the mean age. This validation was performed using New Zealand Census smoking data. For males aged 70 years and over, the prevalence of current cigarette smoking was 10.5%, while the predicted prevalence from smoothing the age-specific curve at the mean age of 77.1 years was 10.0%. The agreement was also reasonable for females, where the average prevalence was 7.6% and the expected was 7.3%.

Where average ages are not available, then life expectancy from the beginning of the age group could be used. However this is expected to be biased for cross-sectional life tables, where the older ages were subject to different mortality experiences to recent mortality rates.

An alternative approach with some merit would be to be use cumulative differencing methods that are popular in demography (Shyrock et al., 1976). This approach would be useful for point estimation, however modelling is expected to become difficult due to correlated errors.

Chapter 8

Estimates for the fitted smoking model

Abstract

Aim: To estimate adjusted transition rates and then to estimate different measures of smoking exposure over the Lexis diagram. **Methods:** The model was fitted backwards in time to estimate adjusted transition rates, and then driven forwards to derive a variety of measures. Model fitting was performed for single-year cohorts born 1920–1965 for males and females. **Results:** Earlier female birth cohorts had lower and later smoking initiation, which influenced other smoking measures. Relative survival for current smokers compared with never and with former smokers was lower among males than females. For males, age-specific duration of smoking and time since cessation varied little between birth cohorts. For females, later birth cohorts had increasing smoking duration and their average smoking duration was less than that for males. Cessation rates for females were imprecise and possibly inconsistent. **Conclusions:** Although there are some concerns with cessation estimates for females, this method provides a variety of useful estimates of smoking exposure.

8.1 Introduction

The main analytical task to this point in the thesis has been the valid estimation of the transition intensities for the smoking model. For the next two chapters, the transition intensities will be used to derive a variety of population-based smoking exposure measures. Several of these measures have not been reported previously.

There were two objectives. The first objective was to describe how the variation in smoking exposure over time and between males and females. The second objective was to provide a useful set of exposure measures for the lung cancer modelling in Chapter 10.

As an outline of the approach, the multi-state model was first fitted backwards in time, starting in 1986 and moving all observed cohorts back to birth. The main purposes of the back fitting were to adjust for differential survival and recall error, and to estimate initiation and cessation rates at younger ages. The back fitting gave a final set of estimates for the transition rates. Using those rates, the multi-state model was pushed forwards from birth to estimate duration parameters.

Measures reported in this chapter include: current smoking prevalence; estimates for cessation and initiation adjusted for all factors; survival ratios between never and current smokers and between former and current smokers; duration of smoking for current and former smokers; and average time since cessation for former smokers. Estimates for dose are reported in Chapter 9.

8.2 Methods

The methods are divided into a section on the back fitting of the model and then a section on the forward modelling.

8.2.1 Back-fitting

The adjustment for differential survival can be estimated by stepping backwards in time from the year of survey. Moreover, this method makes effective use of the oldest reliable prevalence data. For youth, differential survival was considered negligible.

For a given sex and one-year birth cohort c (for 1920–1965 birth years):

1. Set time t as 1986
2. Set age a $(= t - c)$
3. Initialise the state vector
4. Set the recall error rate (0 or 0.02 for females; 0 or 0.01 for males. See Chapter 6)
5. Step through the dynamic model, moving from 1986 to age 10 years
    (a) Calculate prevalence at time t from the state vector

   (b) Extract total mortality rate, rate ratios for current and former smokers, and recalled initiation and cessation rates

      i. Special case: age=25. Fit a dynamic model from age 25 to age 10, estimating a constant for the cessation rate and a scale factor for the initiation rate

      ii. Case: age≤25. Set the cessation rate and initiation rate to the equivalent for that from special case

      iii. Case: calendar year during 1974–1986. Set the cessation rate from ACCV-based data and the initiation rate from RFPS-based data

      iv. Case: calendar year before 1974. Set both the initiation and cessation rates from RFPS-based data

   (c) Calculate never smoker mortality rate

   (d) Calculate survival from the (cumulative) transition probability matrix

   (e) Calculate initiation and cessation rates adjusted for survival

   (f) Calculate the transition intensity matrix

   (g) Calculate the (step) transition probability matrix

   (h) Update the state vector and the cumulative transition probability matrix, and calculate prevalence

   (i) Update year ($= t - 1$) and age ($= a - 1$)

6. Check that estimates for age 10 years are sensible

The back-fitted model provided estimates ending in 1985. For the period 1986–1995, several data sources were used. Cessation and prevalence estimates were taken from the cross-sectional estimates based on the Anti-Cancer Council of Victoria (ACCV) data, as derived in Chapter 7. Initiation rates were taken from smoothed estimates from the Risk Factor Prevalence Study (RFPS; see Chapter 5). This choice was equivocal, as data were also available from the 1998 National Drugs Strategy Household Survey. Both sets of data provided similar results for 1986–1992, however later estimates from the 1998 National Drugs Strategy Household Survey appeared to be inflated, possibly due to recall of recent changes in behaviour.

Estimates derived from the process of back-fitting included prevalence for current and former smokers, and estimates of initiation and cessation rates adjusted for survival and recall error. Unless otherwise noted, estimates were adjusted for recall error.

Modelling for youth

Given the prevalence for never, current and former smokers (represented by π_n, π_c and π_x, respectively) at age 25 years and at age 10 years we want to estimate the initiation rate (α_I) and the cessation rate given the state vector at ages 10 and 25 years, where

$$\pi_n(10) = \exp\left[-\int_0^{10} \alpha_I \, \mathrm{d}u\right]$$

together with $\pi_c = (1-\pi_n)\cdot(1-\epsilon)$ and $\pi_x = \pi_n(1-\epsilon)$. For numerical stability, the constant ϵ was given an arbitrary small value (0.03), representing a small proportion of ever smokers at age 10 who had quit.

It was assumed that cessation was constant across ages 10–24 years (suggested by Model 1C from the census analysis of cessation in youth. See Figure 6.6 on page 168). Alternative forms were investigated, including linear and exponential models, being fixed at age 25 years based on retrospective estimates. However the constant model was the most parsimonious. Smoking misclassification rates for those ages were also assumed approximately the same as for older adults (rate of 0.01 per annum for males and 0.02 per annum for females; no sensitivity analysis for zero recall error was undertaken). Smoking initiation was assumed to follow estimates from recall up to a scale factor as a multiplicative constant.

The average smoking cessation rate and the scale factor for initiation were estimated using least squares. The same model used for youth as for older age groups. For variance estimation, maximum likelihood estimation was found to yield the same point estimates as for the least squares analysis. Initial values used were 0.05 for the average cessation rate and 1.00 for the scale factor.

8.2.2 Forward-fitting

The model was pushed forward in time to calculate other survival estimates and duration estimates. Unless otherwise noted, estimates have been adjusted for recall error.

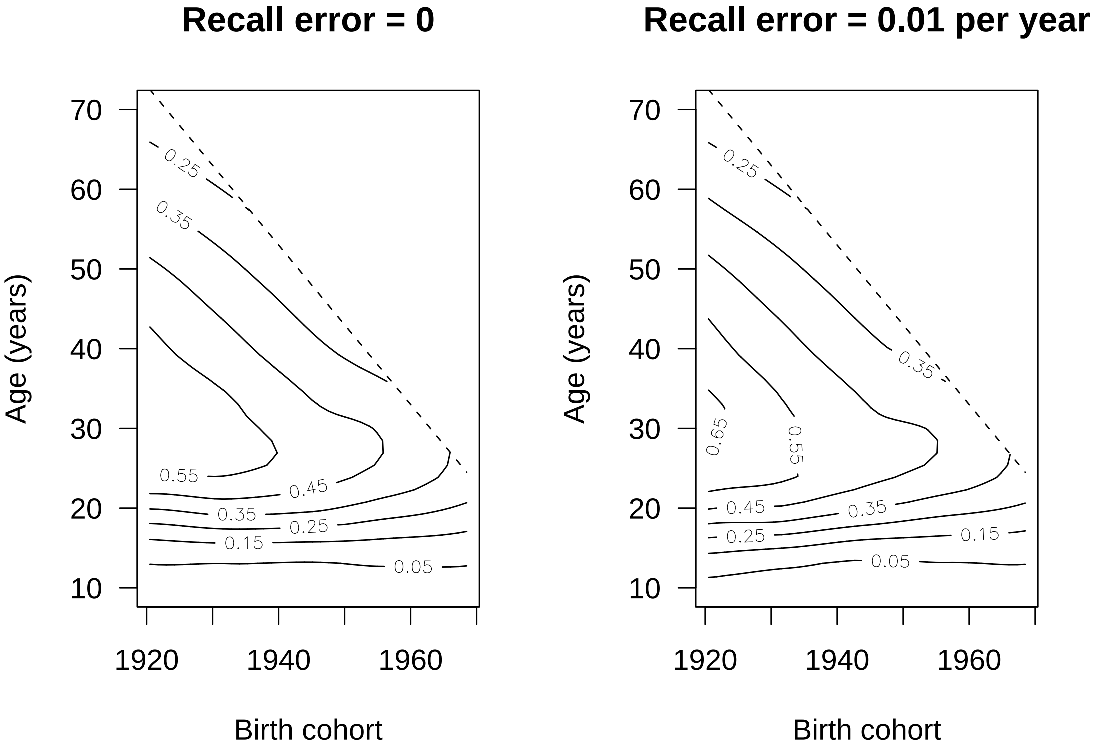
Survival from time s to time t given smoking status at time s was calculated from the transition probability matrices from time s to time t. Relative survival is the ratio of survival for one group compared with the survival for another group. The groups are defined here based on their smoking status at an initial age. These survival estimates are adjusted for cessation and initiation patterns observed up to 1986.

Formulae for duration estimates are derived in Section 3.5.3 on page 63.

8.3 Results

In the figures in this section, the dotted line represents the boundary between where data were and were not available.

Final estimates for current smoking prevalence among Australian males are shown in Figure 8.1. The typical pattern of current smoking prevalence was for a rise at earlier ages, peaking and then a decline at older ages. Peak prevalence occurred in the late twenties or early thirties, with higher peak prevalence at older ages for the earlier cohorts. There was no evidence of stable age-specific prevalence for the earliest cohorts, so that prevalence for the 1900–1919 birth cohorts may have been higher. Finally, recall error had limited influence on the estimates, with estimates adjusted for recall error being slightly higher at younger ages for the earlier cohorts.


Recall error = 0

Recall error = 0.01 per year

Figure 8.1: Estimated prevalence of current smoking, by recall error rate, Australian male cohorts born 1920–1965

The results for females all showed the influence of later uptake by the earliest birth cohorts. In Figure 8.2, the late uptake can be seen by the contour lines being shifted up by age in the earlier cohorts. The level of smoking was considerably less than the level for males. The period decline in smoking prevalence seen among males was also less evident for females. In particular, the younger female cohorts showed evidence for earlier and more sustained smoking uptake. Estimates adjusted for recall error tended to increase prevalence at the younger ages for the earlier cohorts, but otherwise did not affect the other characteristics.

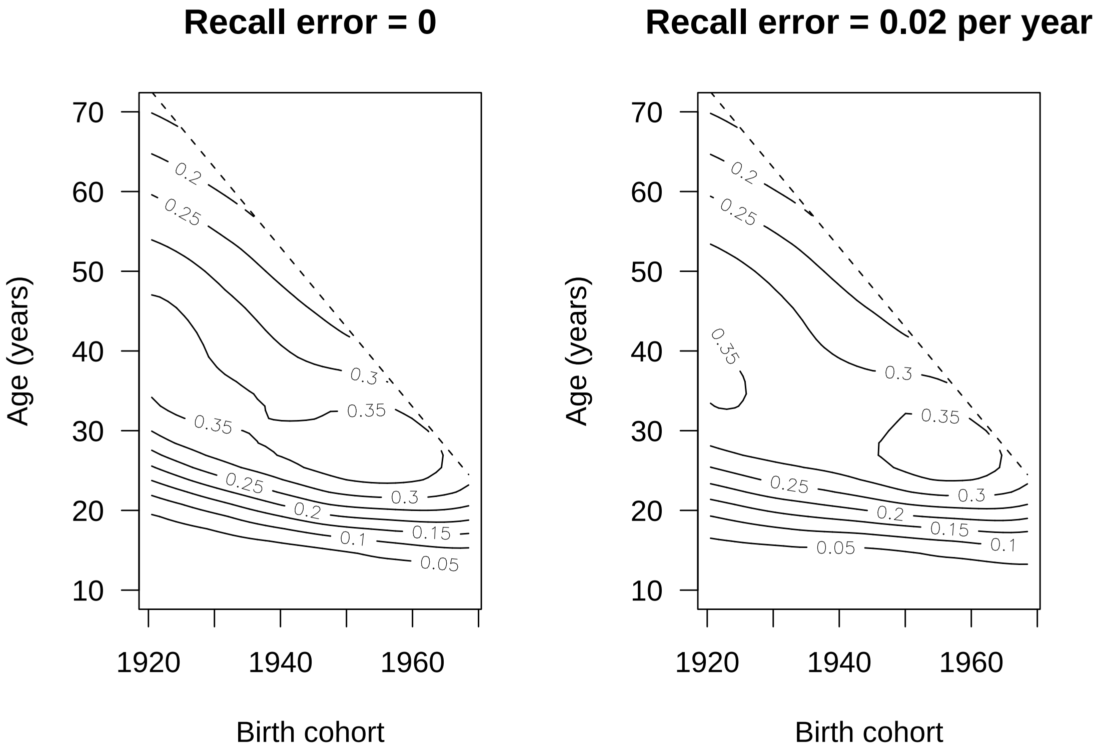


Figure 8.2: Estimated prevalence of current smokers, by recall error rate, Australian female cohorts born 1920–1965

The prevalence of former smokers among males is shown in Figure 8.3. From the recall error-adjusted chart, age-specific prevalence was relatively stable for the earliest cohorts, which would allow modelling for the cohorts born in 1900–1919. The age-specific prevalence of former smokers increased in the later periods, which may be explained by increased smoking cessation. Adjustment for recall error had a more appreciable effect on the prevalence of former smokers compared with that for current smokers. The effect of the adjustment was to increase the estimated prevalence of former smokers for the earlier birth cohorts.

The prevalence of former smoking is influenced by uptake and cessation. There is evidence for both of these effects in the prevalence of former smoking among females (see Figure 8.4). Late uptake in the earlier birth cohorts kept former smoking prevalence low, while increased uptake and cessation led to the broad increase in former smokers in more recent periods. After adjustment for recall error, the 1930–1939

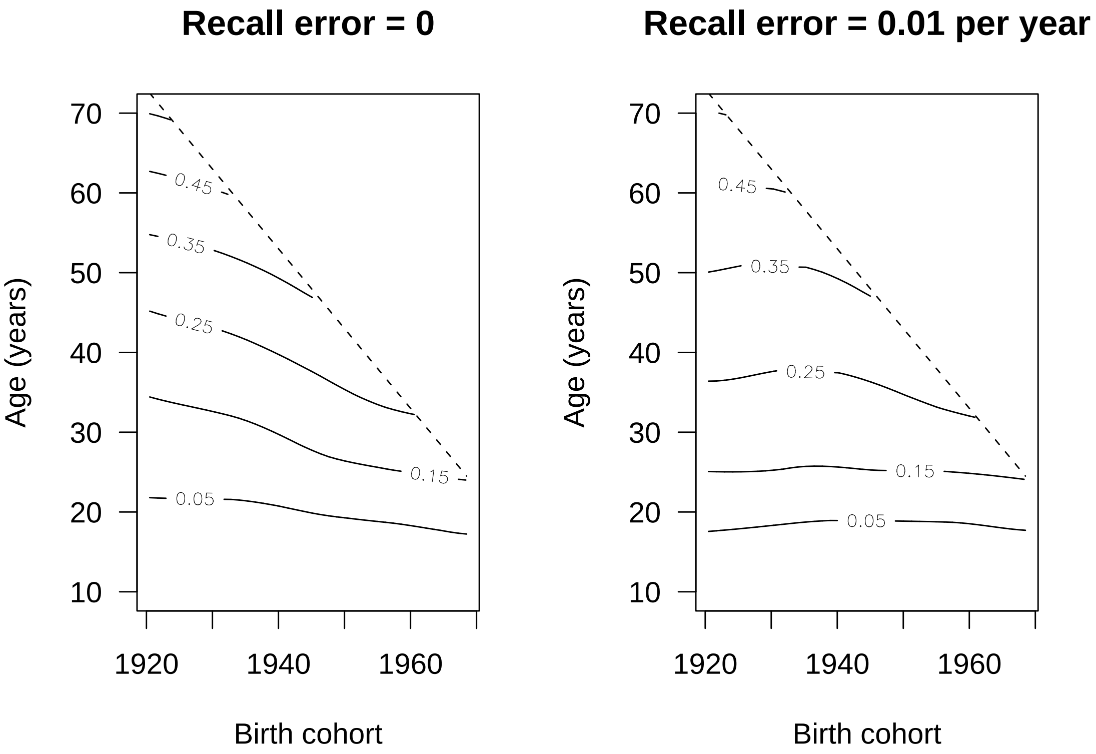


Figure 8.3: Estimated prevalence of former smokers, by recall error rate, Australian male cohorts born 1920–1965

birth cohort had vertical lines for ages 30–39 years, which suggests limited cessation during 1965–1974. Similar to males, the prevalence estimates changed appreciably after they had been adjusted for recall error.

The results for initiation and cessation rates are familiar from Chapters 5–7. Initiation rates adjusted for recall error (0.01 per year for males, 0.02 per year for females) are shown in Figure 8.5. As expected from current smoking prevalence, the initiation rates for earlier male cohorts appear to be increasing. The initiation rate estimates for females suggest that peak initiation was for the birth cohorts born close to 1960. Lower initiation rates among the earlier female birth cohorts is evident. For later birth cohorts, male rates may be in slow decline, while the rates for females may be stable or in very slow decline.

Cessation rates adjusted for recall error are shown in Figure 8.6. The cessation rates for males and females were high at younger ages, dropping at 30–44 years and then increased with increasing age. The increase among males for older ages is considerably greater than that for females. The pattern for females at younger ages is more difficult to interpret. The rise in the teenage years for the 1940s birth cohort could be explained by experimentation, however this is more likely to be a model artefact, where the rapid cessation is due to under-estimated cessation at older ages

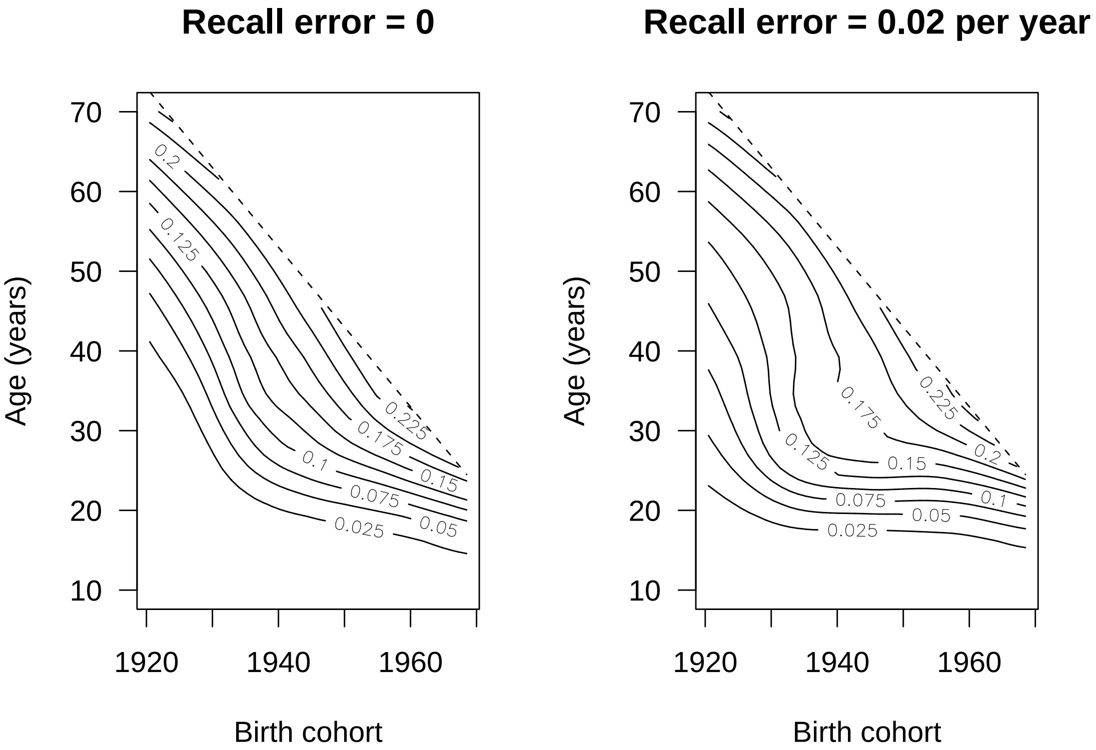


Figure 8.4: Estimated prevalence of former smokers, by recall error rate, Australian female cohorts born 1920–1965

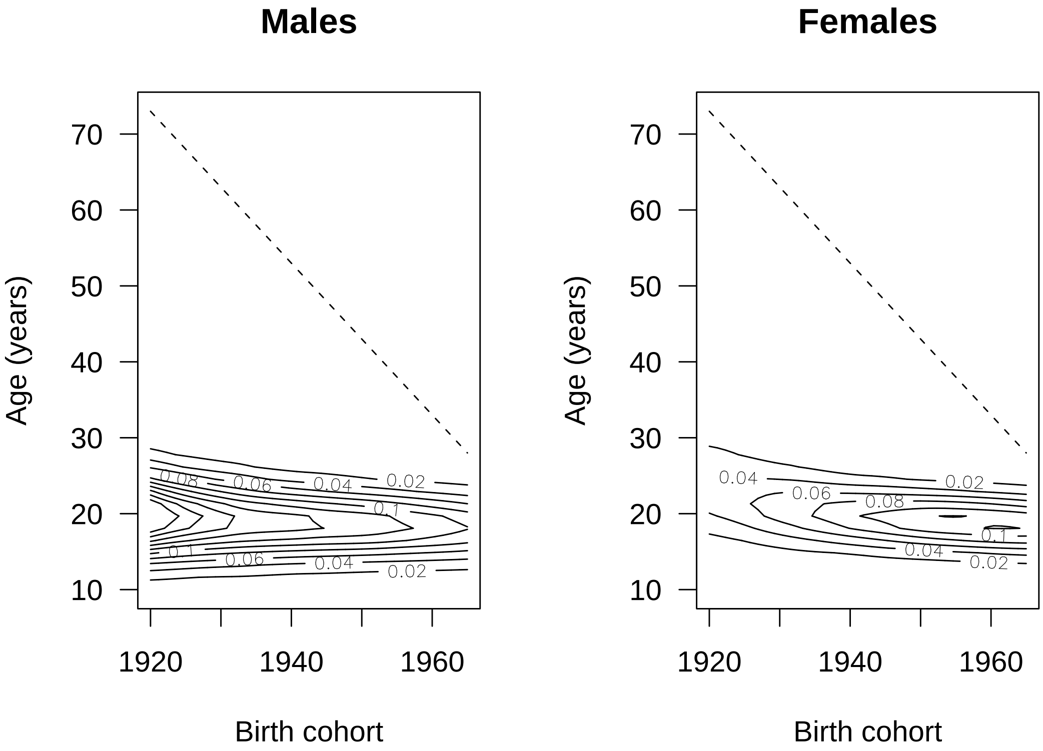


Figure 8.5: Estimated smoking initiation rates per year, Australian male and female cohorts born 1920–1965

for the same cohorts.

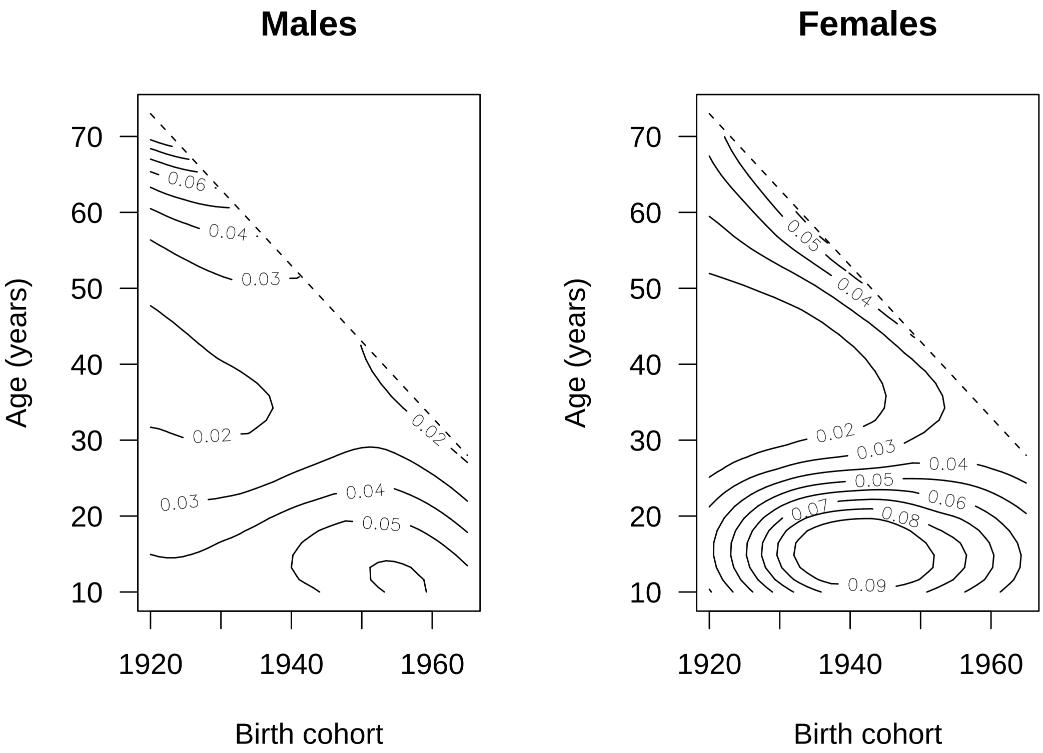


Figure 8.6: Estimated smoking cessation rates per year, Australian male and female cohorts born 1920–1965

Estimates for relative survival for current smokers compared with never smokers are shown in Figure 8.7. Being a current smoker at age 20 years was associated with only a small decrease in survival, possibly because many smokers at age 20 years quit before middle to late life. However by age 30 years, a current smoker had an appreciable decrease in survival, with relative survival decreasing with increasing age.

The estimates of relative survival for males were lower than those for females. This may be explained by the use of the transition probability matrices for 1985 that included females who had later uptake of smoking and lower rate ratios.

Survival for former smokers was intermediate between that for never smokers and for current smokers (Figures 8.7 and 8.8). Aside from sensitivity at the youngest ages, the patterns for the two figures were broadly similar.

8.3.1 Duration estimates

For males, the age-specific duration of smoking for current smokers was uniform across birth cohorts (Figure 8.9). This may not be surprising, given the evidence

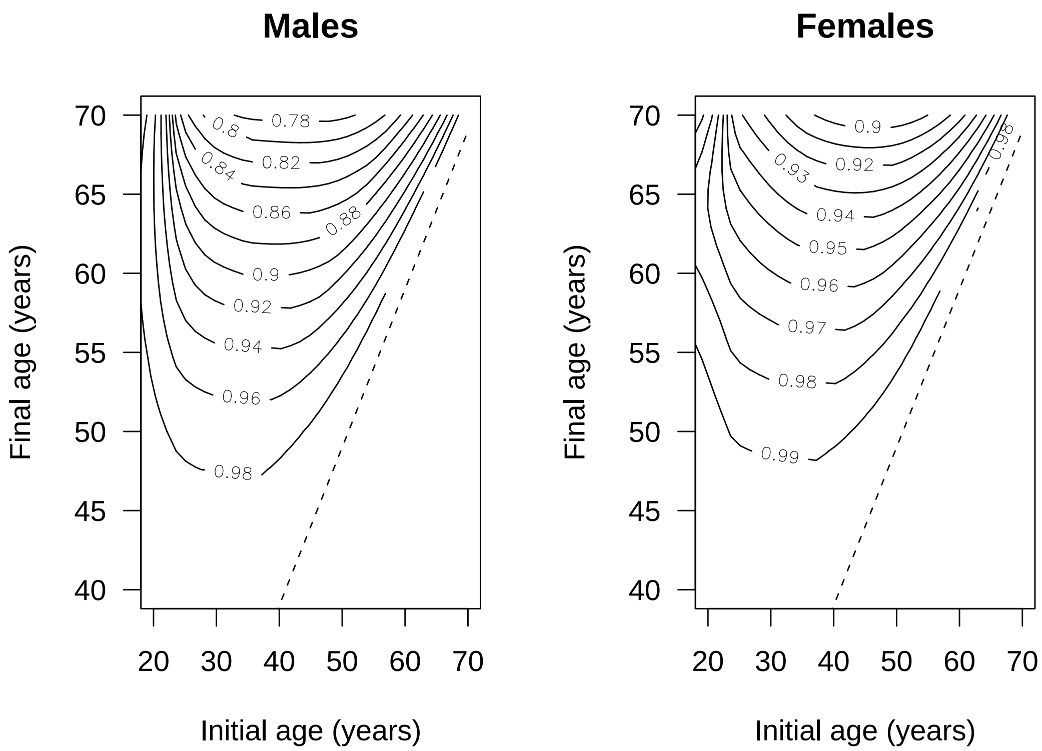


Figure 8.7: Relative survival for current smokers at the initial age compared with never smokers at the initial age, by sex (1985)

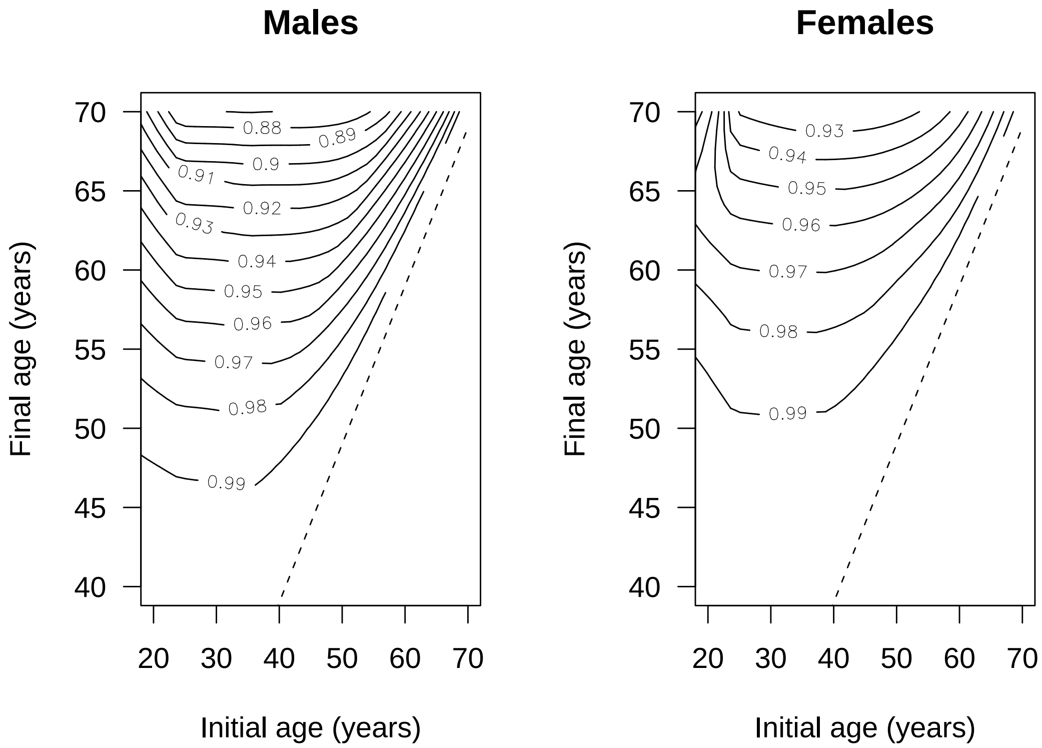


Figure 8.8: Relative survival for current smokers at the initial age compared with former smokers at the initial age, by sex (1985)

that the shape for smoking initiation had not changed appreciably across male birth cohorts. Among females, the duration of smoking for current smokers for a given age increased considerably for later birth cohorts, such that female smoking duration was only slightly lower than male smoking duration for the most recent period. The differences in duration between males and females can be explained by uptake at a later age among the female birth cohorts, with latest uptake seen in the earliest female birth cohorts. In summary, female current smokers smoked for fewer compared with males, however the differences between the sexes were small for the most recent period.

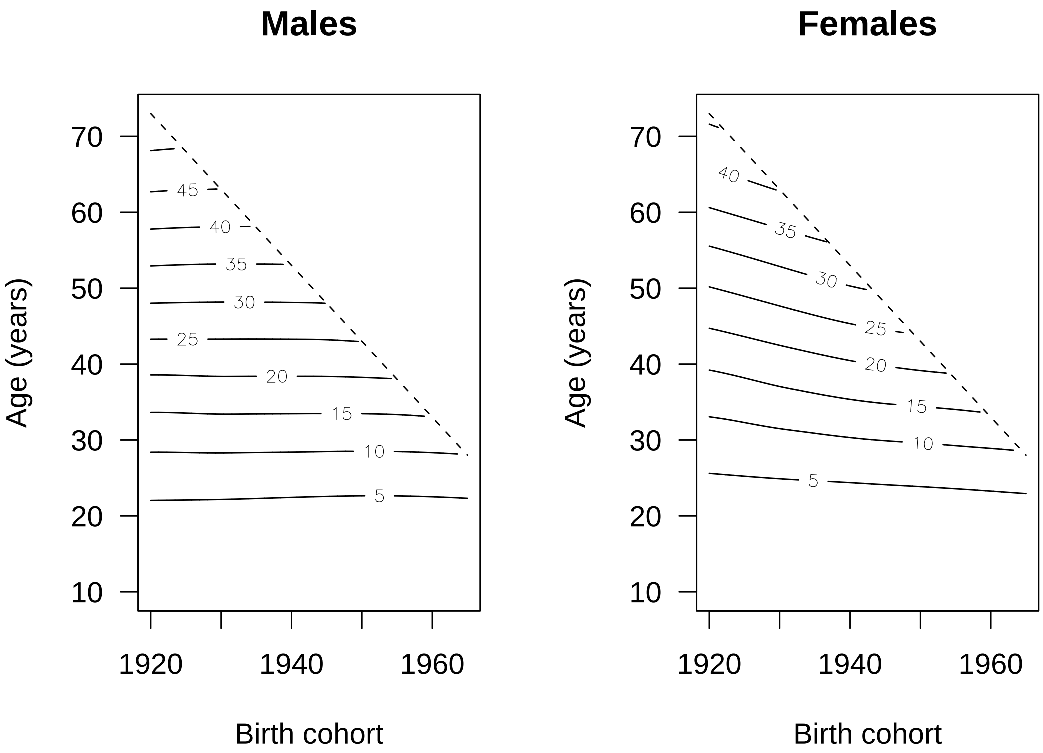


Figure 8.9: Average duration of smoking for current smokers, Australian male and female cohorts born 1920–1965

For the average duration of current smoking by former smokers, males again had a stable set of age-specific estimates between birth cohorts (see Figure 8.10). The age-specific estimates for females were slightly lower than the rates for males. It is unclear whether the decreased duration of smoking for females as suggested by the convex contours is a real period effect or a model artefact.

The average time since cessation for former smokers is shown in Figure 8.11. The age-specific estimates for males were stable at younger ages. There is a suggestion of change at older ages, however this may be affected by imprecise estimates of rapid cessation at older ages (see also the oldest ages for males in Figure 8.10). The

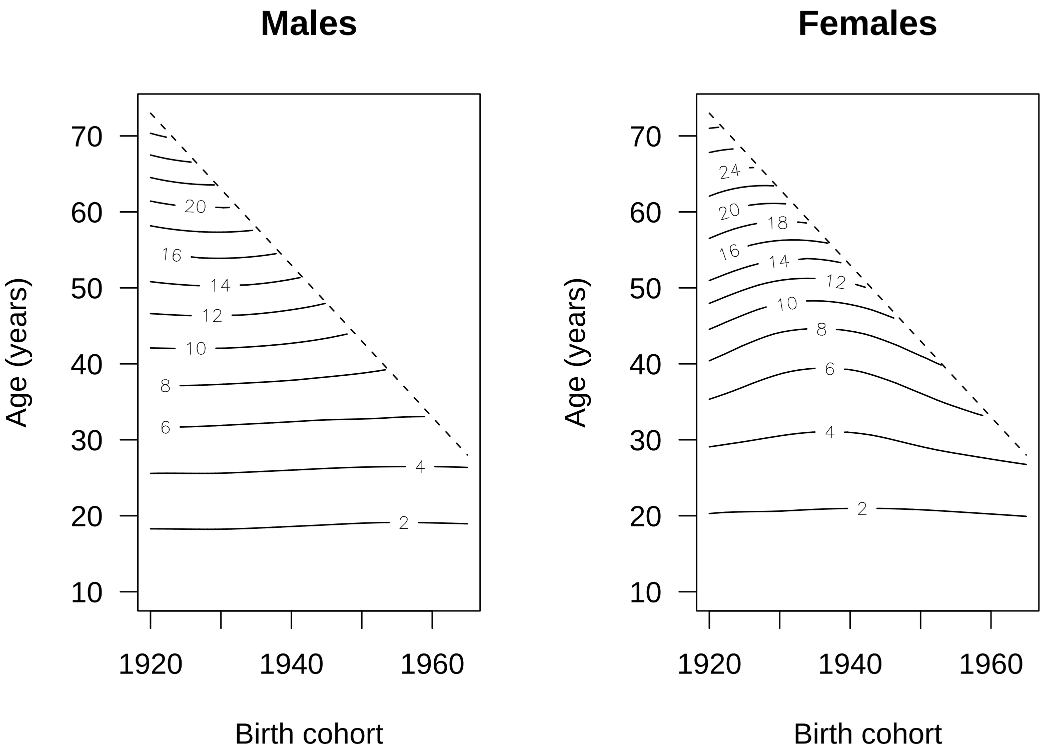


Figure 8.10: Average duration of smoking for former smokers, Australian male and female cohorts born 1920–1965

pattern for females was for a slower rise in time since cessation for the earliest birth cohorts, which can be interpreted in the context of late smoking uptake and hence later smoking cessation.

8.4 Discussion

In summary, a method is presented to adjust the transition rates and to estimate a variety of smoking measures. Earlier female birth cohorts had lower and later smoking initiation, which influenced other smoking measures. Relative survival ratios for current smokers compared with never and with former smokers were higher among males than females. For males, the age-specific duration of smoking and time since cessation varied little between birth cohorts. Females had increasing smoking duration for later birth cohorts, but their average smoking duration was less than that for males.

The validity of these estimates is dependent upon the validity and precision of the estimated transition intensities. The main efforts in Chapters 5–7 were to provide the best possible estimates. However some of the estimates are not consistent with *a priori* expectations. In particular, the cessation pattern during teenage

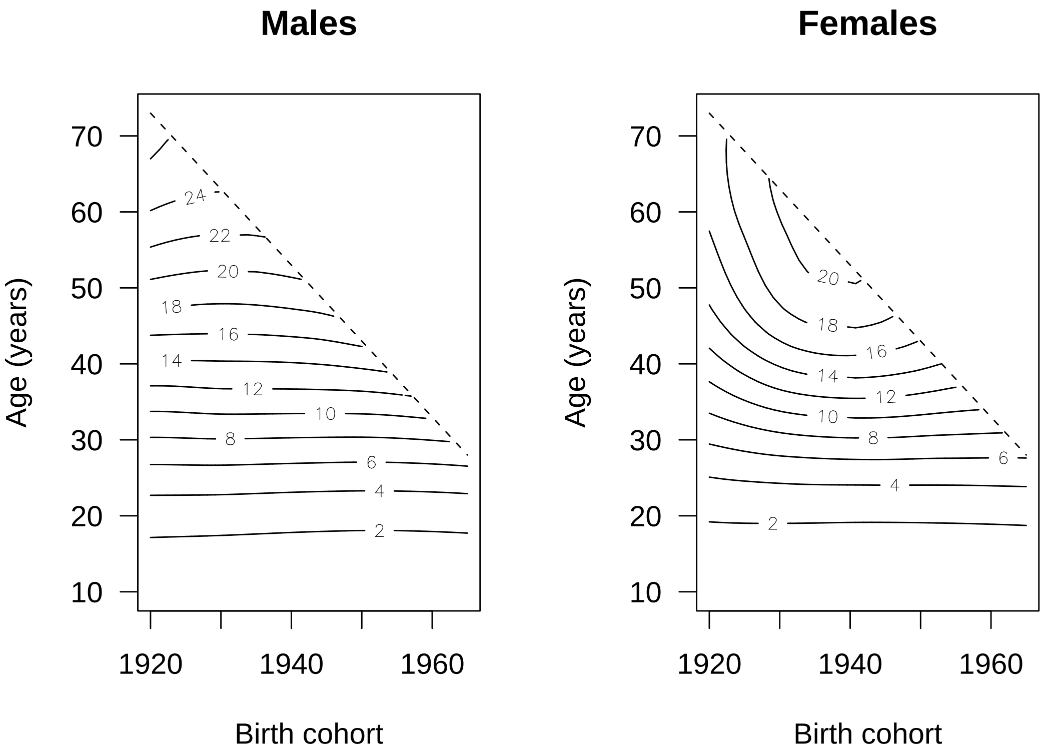


Figure 8.11: Average duration of smoking cessation for former smokers, Australian male and female cohorts born 1920–1965

years for females born 1940–1949 looks unusual. If such a pattern is due to underestimated cessation at later ages, then duration estimates may be biased upwards, which may explain the convex curves in duration intervals for female former smokers (Figure 8.10).

Relatively light assumptions were made in the modelling, except for the choice of estimates for recall error rates. The sensitivity analysis for recall error suggests that the prevalence of former smokers is sensitive to the estimate of recall error, while the prevalence of current smokers is less sensitive to the estimate of recall error. Any modelling that requires good estimates for the prevalence of former smokers should consider the inclusion of this factor. Methods to obtain estimates of recall error rates for different populations require further investigation.

Challenges with estimating past changes in smoking behaviour make future predictions difficult. Although the *rates* of change are difficult to predict, the *cumulative measures* tend to be more stable, particularly for males. Predicted measures for females are likely to be the least precise given the marked changes between cohorts. A heuristic approach would be to assume that the patterns for females become more like males, as observed for a number of measures.

Aetiological studies have noted the variation in dose and duration between sexes and between ages (Burns et al., 1997a). Population-based changes in prevalence of current and former smokers have been presented by age and cohort (e.g. Burns et al., 1997b), however Australian-specific results have not been published. To my knowledge, there are no previous publications of population-based estimates of smoking duration over the Lexis diagram.

Results for changes in duration over the Lexis diagram provide some explanation of the variations in lung cancer mortality between sexes, between cohorts and over time. Duration of smoking by females has typically been less than that for males, which explains why the risk ratios for females have been less. Of greater concern, for more recent periods the smoking duration for females is more similar to males, so that the average lung cancer risk per female smoker is likely to higher.

Before investigating quantitative relationships between smoking duration and lung cancer, consumption data will be applied to the smoking model to estimate dose (Chapter 9).

Chapter 9

Smoking dose estimation

Abstract

Aim: To estimate different measures of smoking dose over the Lexis diagram. **Methods:** Youth smoking dose was estimated using Australian youth smoking surveys. Adult dose estimates were compared between surveys. The 1976 New Zealand Census was used as the reference dose distribution and was combined with population and prevalence estimates to calculate a dose scale factor. The dose estimates were included in the smoking model to derive different measures. **Results:** Daily cigarette consumption for current smokers varied by age, between sexes and over time. For data from 1976–1983, dose was relatively stable between surveys. Dose peaked in middle age and was higher for males than females. Dose was highest during 1960–1975, with dose lower prior to and after that period. For the most recent periods, cumulative measures of age-specific dose were stable or dropping for both current and former smokers. **Discussion:** Several strong assumptions were made about the dose distribution in the absence of better data. This method provides a variety of useful population-based dose estimates.

9.1 Introduction

Measures of smoking duration were estimated in Chapter 8 by applying fitted estimates of smoking transitions to the smoking model. This chapter investigates methods to estimate different measures of smoking dose.

As described in Chapter 1, smoking dose is closely related to risk of lung cancer. After adjustment for smoking duration, daily cigarette consumption has a linear to quadratic association with lung cancer rate (Doll and Peto, 1978).

For the purposes of modelling lung cancer mortality, dose would ideally measure the level that target tissues are exposed to the carcinogenic agent. This poses a difficulty for smoking, as there have been substantial changes in the construction of cigarettes (Wilkenfeld et al., 2000; Burns et al., 2001). It is unclear how to measure the effect due to factors such as the use of filters, reduction in machine-measured tar and changing inhalation patterns. A simple approach is to use cigarette consumption as a metric. This allows the use of results from a variety of aetiological studies that have used a similar metric. Moreover, recent aetiological studies have not observed a decline in lung cancer risk as predicted by ostensibly "lower tar" cigarettes, suggesting that adjusting for machine-estimated tar may bias results. Dose estimates have not been adjusted for tar levels in this chapter. This aspect will be considered further in Chapter 10.

"Dose" can be represented by a variety of measures. Measures may consider a combination of: consumption per capita or per smoker; daily dose or cumulative dose; dose for current or former smokers; daily dose for a given period or average daily dose for a person's smoking history. Moreover, it would be useful to also consider measures of variation of dose in a population, rather than restricting our attention to average dose in a population group. Following Section 3.5.4, *dose* will be defined as the rate of cigarette consumption per smoker, usually expressed as the number of cigarettes (or equivalent) per day.

There is mixed evidence that dose has been changing over time. Self-reported dose from CPS-I and CPS-II showed that daily cigarette consumption by smokers was considerably higher in the 1980s than the 1960s, that daily consumption was higher among males than females, and that daily consumption varied by age, being lower at older and younger ages (Burns et al., 1997b). For Australia, average self-reported daily consumption data are available since 1980, with no suggestion of any period trend (Hill et al., 1990, 1993; Hill and White, 1995; Hill et al., 1999).

One potential issue is that the self-reported data may be biased (see Warner, 1978). Although Jackson and Beaglehole (1985) used New Zealand census data to suggest that any bias may be consistent over time, the use of tobacco consumption data from Customs and Excise may provide a more objective measure.

As an outline of my approach to estimate smoking dose, youth estimates for smoking dose were first calculated. Adult age-specific dose estimates based on self-report from New Zealand censuses and Australian surveys and from CPS-I and CPS-II were then compared to see whether age-specific dose ratios were constant over time. Then the dose estimates from the 1976 Census and for youth were combined with approximate estimates of current smoking prevalence for 1950–1999 and observed consumption to estimate a dose scale factor.

Smoking dose per current smoker was then well defined over the Lexis diagram. Using estimates of dose in the smoking model, various measures of dose and consumption were derived.

9.2 Methods

9.2.1 Youth smoking estimates

Data on youth smoking were available from the Australian Cancer Society secondary students smoking and alcohol surveys for 1987 to 1996 (Hill et al., 1990, 1993, 1995, 1999).

Estimates were required for the smoking dose and current smoking prevalence among youth aged 10–14 years. Moreover, dose by single year of age was considered useful for careful cumulative dose estimation for each birth cohort.

For youth, estimates for dose were only available for respondents who had smoked in the last week. Information was also available on the proportion of respondents who had smoked in the last week and the proportion who had smoked on six or more days in the past week. The latter definition was assumed similar to the “daily”, “regular” or “current” smoking definitions used for adults. Data were only available for those aged 12–17 years.

To estimate dose for current smoking from dose for those who smoked in the last week, an upper bound can be found by assuming that the dose for weekly smokers who are smoking less than daily is negligible. As the daily smokers are assumed to smoke all of the tobacco products, the daily dose is bounded by the weekly dose divided by seven times the ratio of proportions of weekly to daily smokers:

$$\begin{aligned}\text{Daily dose for daily smokers} &< (\text{Weekly dose for weekly smokers})/7\\ &\quad \times \frac{\text{Proportion of weekly smokers}}{\text{Proportion of daily smokers}}.\end{aligned}$$

For a reasonable point estimate, we can arbitrarily assume that dose takes an exponential distribution, where most young weekly smokers are expected to smoke relatively few cigarettes per week. As an extension, the choice of a parametric distribution for smoking dose could be investigated using survey data. With these assumptions, an estimate of the daily dose for daily smokers is

$$\begin{aligned}\text{Daily dose for daily smokers} &\approx \text{(Weekly dose for weekly smokers)}/7 \\ &\quad \times \left[1 - \log \frac{\text{Proportion of daily smokers}}{\text{Proportion of weekly smokers}}\right].\end{aligned}$$

For estimates for those aged 10-14 years, prevalence for those aged 10-11 years was assumed negligible. Therefore the prevalence for those aged 10-14 years was the sum of the prevalence for individual years for ages 12-14 years divided by 5. Dose for those aged 10 and 11 years were estimated by linear extrapolation from those aged 12 and 13 years.

Dose and prevalence for individual years and for ages 10-14 years were averaged across the secondary school surveys. In the absence of other data, youth smoking results for New Zealand were assumed to be broadly similar to Australia.

9.2.2 Adult dose estimates

Data sources

Data on the number of cigarettes smoked daily were available from the New Zealand Census for Population and Dwellings from 1976 and 1981 (Department of Statistics, 1979, 1983). Data were provided as the number of respondents by five year age groups (15–19, . . . , 75 and over) by the number of cigarettes smoked (under 5, 5–9, . . . , 50 and over) by sex. The cumulative curves by dose were reasonably smooth, suggesting that the mid-points for each dose category could be used. The open dose category of 50 cigarettes and over used an upper bound of 70 cigarettes per day and a mid-point of 60 cigarettes per day. The estimated mean dose was insensitive to the value of the mid-point for the open dose category. Means and variances for dose by age group and sex were estimated using the mid-points.

Results for daily cigarette consumption from the 1980 ACCV smoking survey were available from Hill and Gray (1982). The population standard deviation for dose for an age group was estimated from the mean dose times the average coefficient of variation (=0.58) from the 1976 New Zealand Census. The variance of the mean dose was estimated by the population variance for dose divided by the number of smokers times a design effect of two.

Similar estimates from the 1983 and 1989 surveys for the Risk Factor Prevalence Study were estimated from an electronic data file by weighting individual responses. The questions were “I currently smoke $\underline{\text{XX}}$ manufactured cigarettes a day” and “I currently smoke $\underline{\text{XXX}}$ grams 'hand-rolled' cigarette tobacco per week”.

It was assumed that one gram of hand-rolled cigarette tobacco was equivalent to one manufactured cigarette. Estimates were weighted by the sample weights. The approximate variance of the mean dose for an age group was estimated by scaling the sample weights to sum to the sample count.

The average change in daily cigarette consumption rate was estimated from the first Cancer Prevention Study (CPS-I) in the 1960s to the second Cancer Prevention Study (CPS-II) in the 1980s using published data (Burns et al., 1997b). Data were available by five-year age groups for mean dose and 95% confidence intervals.

Tobacco products available for consumption from Excise and Customs and Australian population estimates were obtained from the Australian Bureau of Statistics. All data sources are described in detail in Chapter 2.

Comparison between data sources

The two censuses from New Zealand had large numbers, so that a graphical comparison was sufficient to assess whether age- and sex-specific dose varied over time.

A generalised linear model was used to compare the 1976 New Zealand Census with the three Australian data sources by sex. The observations were weighted by their inverse variances. The model assumed normal errors and a log link. The model included dose as an outcome, with a quintic polynomial in age, survey as a factor variable, and an interaction between survey and linear age.

For the comparison between CPS-I and CPS-II, confidence intervals were used to estimate the variance of the mean estimates. A generalised linear model was fitted using a log link with the inverse variance as weights (McCullagh and Nelder, 1989). The models included main effect for study together with quadratic age terms for females and cubic age terms for males. As a cross-check, generalised additive models were fitted using spline terms for age (Hastie and Tibshirani, 1990).

9.2.3 Changes in dose during 1950–2010

Observed tobacco consumption for Australia was compared with expected consumption based on a reference distribution with adjustment for the estimated number of smokers. Observed data were taken from Customs and Excise data (see Chapter 2).

Expected consumption estimates were calculated from a reference distribution of mean daily dose, expressed in cigarette equivalents per day, multiplied by the number of current smokers multiplied by 365. The 1976 New Zealand Census provided the reference dose distribution. Estimates for the prevalence of current smoking for 1920–1970 birth cohorts were available from Chapter 8. A strong assumption was made that earlier cohorts were stable at the level of the 1920 cohorts. With this assumption, approximate numbers of smokers were calculated for 1950–1995 using prevalence and population estimates.

By assuming that the age and sex distribution of dose was known up to a constant over time, the ratio of observed divided by expected dose gave the scale factor for the 1976 New Zealand Census dose distribution. If $\text{Cons}_{\text{Customs \& Excise}}(t)$ is cigarettes available for consumption for calendar year t, $\text{Dose}_{\text{Census}}(a)$ is daily dose of cigarettes from the 1976 New Zealand Census, $\text{Pop}(a, t)$ is the population aged a at time t, and $\hat{\pi_c}(a, t)$ is estimated smoking prevalence, then the adjustment factor K is

$$K(t) = \frac{\text{Cons}_{\text{Customs \& Excise}}(t)}{365 \sum_a \text{Dose}_{\text{Census}}(a)\hat{\pi_c}(a,t)\text{Pop}(a,t)}.$$

The dose for current smokers was then calculated over the Lexis diagram by multiplying the dose distribution from the 1976 Census by the adjustment factor K:

$$\hat{\text{Dose}}(a,t) = \text{Dose}_{\text{Census}}(a)K(t).$$

Values were smoothed using local likelihood with a log link and normal errors, and then extrapolated back to 1940 and forward to 2000. In the absence of better information, estimates prior to 1940 and following the year 2000 were assumed approximately constant.

9.2.4 Forward fitting of dose parameters

Dose can then be expressed as daily consumption per capita and per current smoker. By applying the dose data to the smoking model, a variety of other dose measures were calculated over the Lexis diagram. These included: average dose across the smoking history for current and former smokers; and cumulative cigarette consumption per capita, per current smoker and per former smoker. Cumulative consumption is expressed in terms of *pack-years*, where one pack-year is defined as smoking 20 cigarettes (or equivalent) per day for one year (7300 cigarettes or equivalent).

See Section 3.5.4 on page 65 for details.

Age (years)	Males	Females
10	1.2	0.6
11	2.6	2.1
12	3.9	3.5
13	5.2	4.9
14	6.8	6.3
15	8.0	7.1
16	8.9	8.1
17	9.7	8.2
10–14	6.1	5.7

Table 9.1: Estimated daily cigarette consumption for daily smokers, Australian secondary students

9.3 Results

9.3.1 Youth smoking estimates

The estimated prevalence of daily smokers aged 10–14 years was comparatively stable across the surveys before 1996, with a suggestion of an increase in 1996. For parsimony, the average for the surveys was used. For Australian secondary students, approximately 2.2% of boys and 2.0% of girls aged 10–14 years were estimated to be daily smokers during the period 1987–1996.

Dose estimates for Australian secondary students are shown in Table 9.1.

9.3.2 Adult dose estimates

Comparison between data sources

For a comparison of dose between the 1976 and 1981 New Zealand censuses, the 1980 ACCV survey and the 1983 and 1989 surveys from the Risk Factor Prevalence Study, see Figure 9.1. Daily cigarette consumption was generally higher among males than females, being lowest at the younger and older ages. Peak dose in most surveys was at 40–50 years of age. The two New Zealand censuses provided quite similar results, suggesting that there was little change in behaviour between 1976 and 1981. The Australian sample surveys had considerably more sampling variation.

A formal comparison between the 1976 New Zealand Census, the 1980 ACCV survey and the 1983 Risk Factor Prevalence Study survey suggests that the dose estimates were not significantly different. In contrast, dose estimates from the 1989 Risk Factor Prevalence Study had a steeper slope by age ($RR - 1 = 0.0077$, 95% confidence interval: 0.0004, 0.0151 for males; $RR - 1 = 0.0103$, 95% confidence

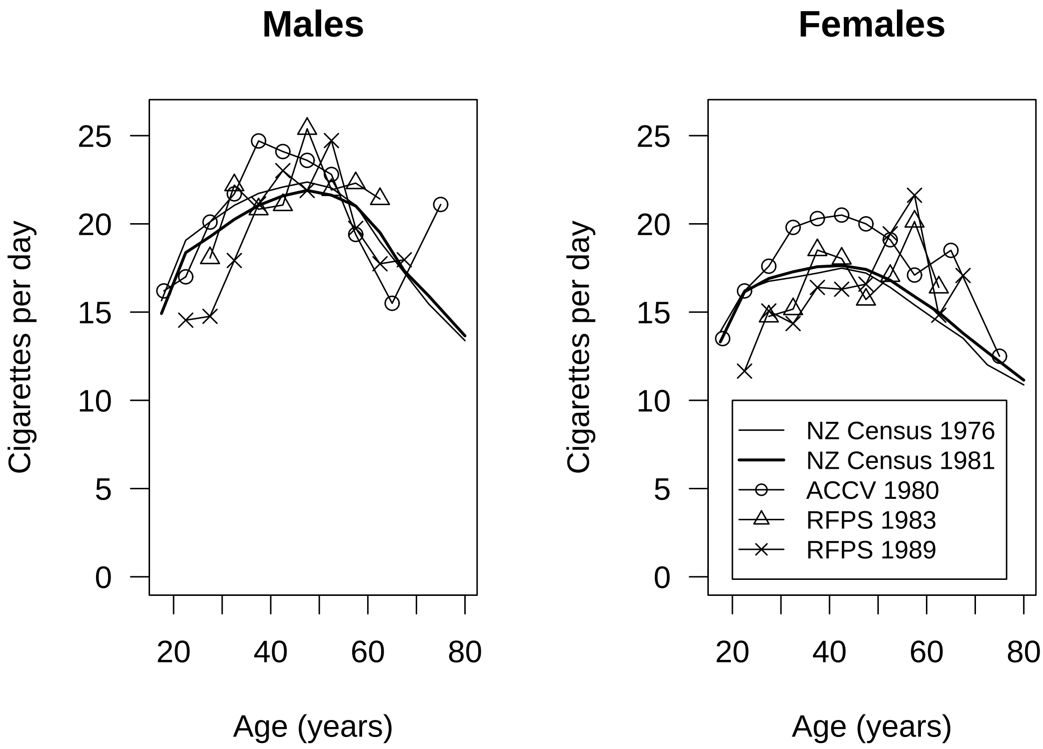


Figure 9.1: Age-specific daily cigarette consumption from different Australian and New Zealand surveys, by sex

interval: 0.0012, 0.0194 for females).

For a comparison between CPS-I and CPS-II, the average increase in consumption rates going from CPS-I to CPS-II was 13.8% (95% confidence interval: 11.6, 16.0) for males and 27.4% (95% confidence interval: 24.3, 30.5) for females. Dose estimates from CPS-II were steeper by age compared with estimates from CPS-I, although the level of change per year of age between the two studies was small ($RR - 1 = 0.37$, 95% confidence interval: 0.17, 0.57 for males; $RR - 1 = 0.19$, 95% confidence interval: 0.05, 0.33 for females).

In the absence of good historical information on changes in dose ratios, it was assumed that dose estimates from the 1976 New Zealand Census could be used as a standard distribution for Australia and New Zealand. Dose estimates from the 1976 Census, including the mean and the population standard deviation, are shown in Table 9.2. The average coefficient of variation was 58%, with moderately large standard deviations. There was limited variation in the coefficient of variation between age groups and between males and females.

Age group	Males		Females	
(years)	Mean	SD	Mean	SD
15-19	15.6	10.2	13.9	9.2
20-24	19.1	10.8	16.2	9.2
25-29	20.1	10.7	16.7	9.2
30-34	21.0	10.9	17.0	9.2
35-39	21.7	11.4	17.2	9.4
40-44	22.1	11.7	17.5	9.5
45-49	22.4	11.7	17.2	9.4
50-54	22.0	11.7	16.4	9.0
55-59	21.1	11.5	15.4	8.6
60-64	19.0	10.7	14.4	8.3
65-69	17.2	10.0	13.5	8.4
70-74	15.5	9.4	12.0	7.7
80+	13.4	9.1	10.9	8.2

Table 9.2: Daily cigarette consumption per smoker, 1976 Census of Population and Dwellings

Calendar year	Males	Females
1950	0.538	0.218
1955	0.517	0.241
1960	0.484	0.253
1965	0.446	0.255
1970	0.416	0.259
1975	0.390	0.266

Table 9.3: Estimated prevalence of current smoking in Australia, by sex (using prevalence estimates for 1920–1970 birth cohorts and assuming earlier cohorts had prevalence similar to the 1920 cohort)

9.3.3 Changes in dose during 1950–2010

The estimated prevalence of current smoking in Australia for males and females for 1950–1975 are shown in Table 9.3. The 1950 estimate for male smoking prevalence of 54% was at variance with published prevalence estimates of 72% in 1945 (Woodward, 1984) and of 69% in 1950 (Tyrrell, 1999). The estimated prevalence for female smoking in 1950 of 22% was also lower than expected from Woodward (1984) and (Tyrrell, 1999) (26% in 1945 and 27.5% in 1950, respectively).

Expected consumption was calculated by multiplying estimated populations by smoking prevalence by the 1976 New Zealand Census dose estimate by 365 across age groups and sex, and then by aggregating for year. The ratio of observed consumption divided by expected consumption is shown in Figure 9.2. Dose changed appreciably over time. There is strong evidence for a rapid decline in dose since the mid 1970s. The two data points prior to 1960 were considerably lower than the results for 1960–1975. Again there is no evidence for a plateau in the earlier results, so that dose estimates for years prior to 1950 become increasingly imprecise.

Values for the smoothed function are shown in Table 9.4.

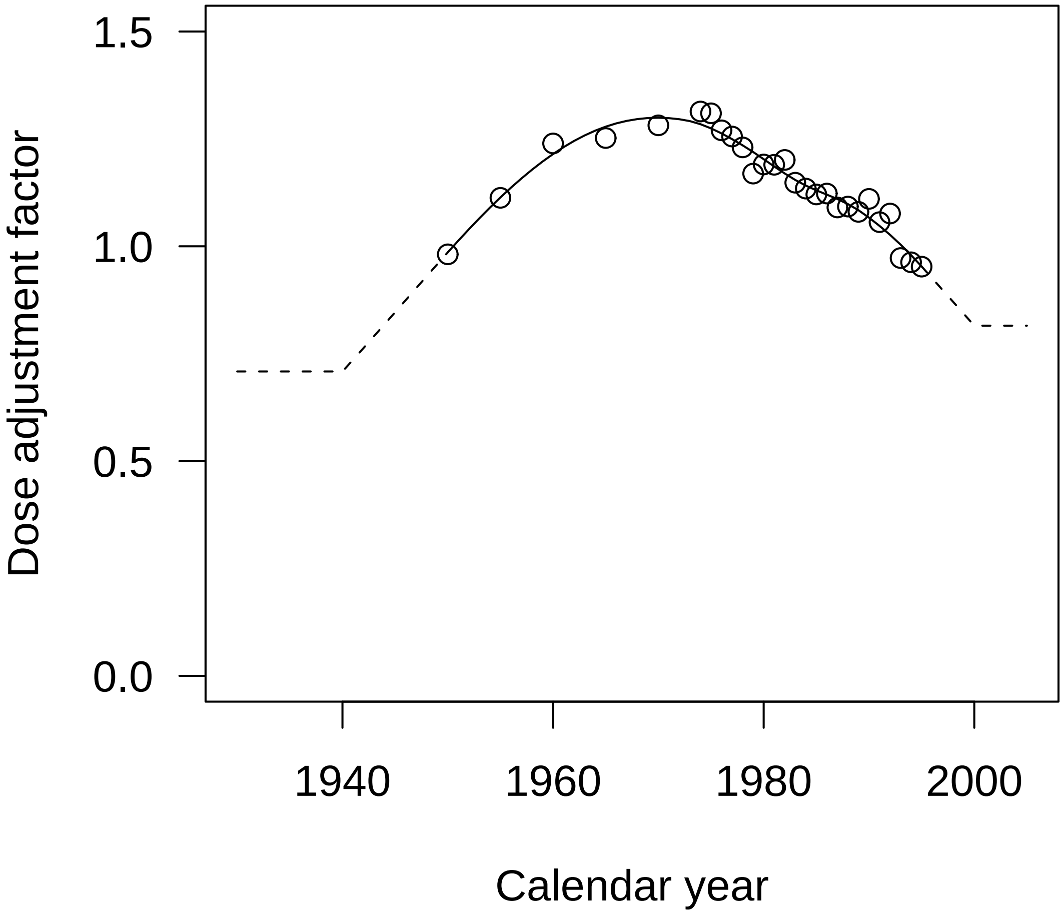


Figure 9.2: Adjustment of dose relative to the 1976 New Zealand Census

9.3.4 Forward-fitting for dose parameters

Dose estimates expressed as daily cigarette consumption per smoker are presented in Figure 9.3. The period pattern is a consequence of the estimates being the product of dose estimates for the 1976 New Zealand Census multiplied by the common adjustment factor in Table 9.4. Any differences between sexes will be an expression of differences from the Census results. This age-period model for dose may not be valid, particularly for the earlier periods, where the distribution by age and sex may vary. All of the later estimates in this chapter are dependent upon the validity of

Calendar period	Adjustment factor
1930	0.46
1935	0.58
1940	0.71
1945	0.85
1950	0.99
1955	1.11
1960	1.21
1965	1.28
1970	1.30
1975	1.27
1980	1.20
1985	1.13
1990	1.07
1995	0.95
2000	0.82

Table 9.4: Estimated dose adjustment factors relative to the 1976 New Zealand Census, Australia

these estimates.

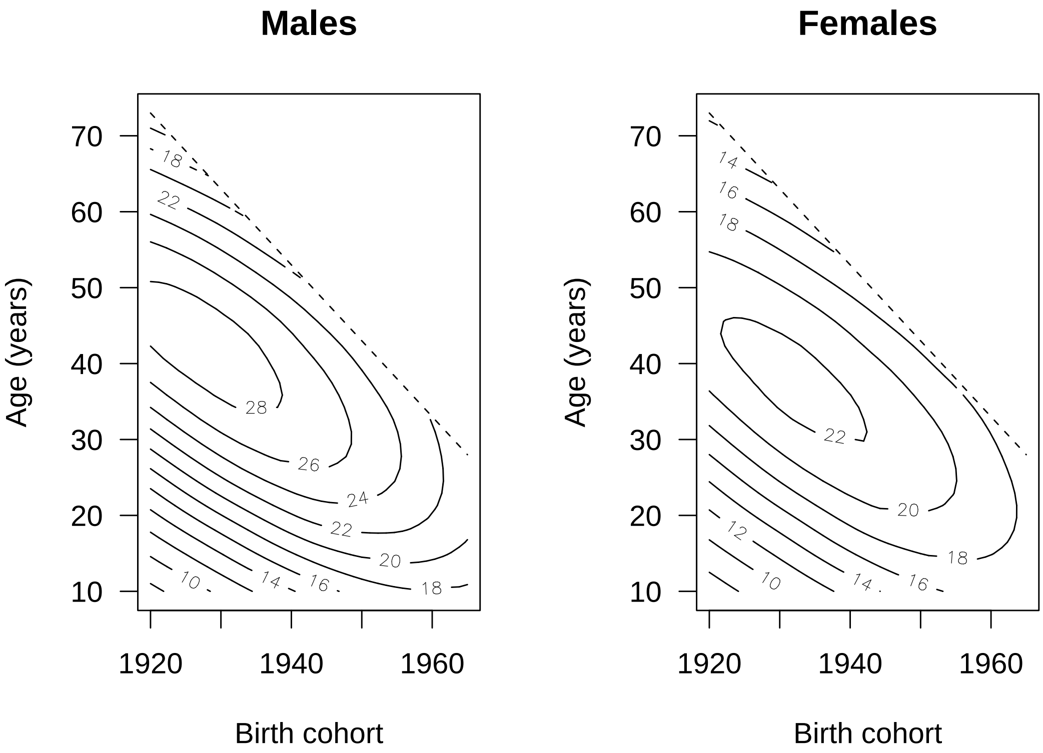


Figure 9.3: Estimated dose (cigarettes per day) per current smoker, male and female cohorts born 1920–1970

The pattern for daily cigarette consumption per capita (Figure 9.4) was very similar to the pattern for current smoking prevalence (Figures 8.1 and 8.2). This is not surprising, given that the level of variation in dose was limited and per capita consumption was the product of daily consumption per smoker times smoking prevalence.

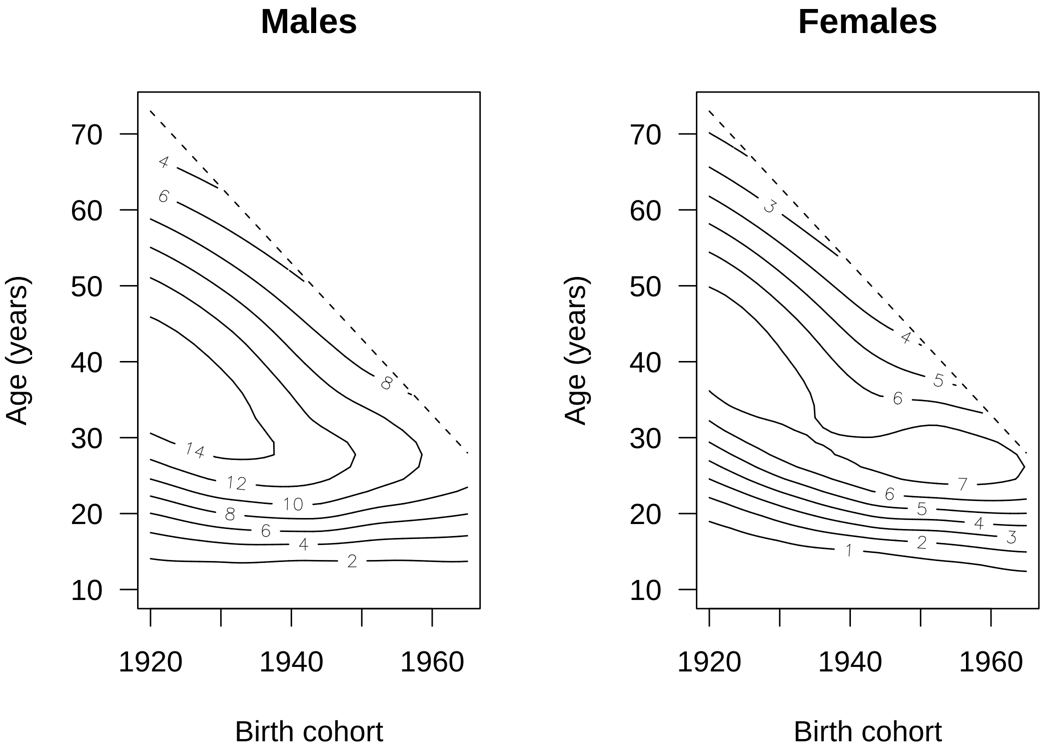


Figure 9.4: Estimated consumption (cigarettes per day) per capita, male and female cohorts born 1920–1970

The pattern for average daily dose since smoking initiation for a current smoker, adjusted for cessation and differential survival, was different from daily dose at a given time (see Figures 9.3 and 9.5). The peak average dose was at a later period, with a greater spread as a result of averaging. The pattern (or shape) was again similar in males and females, although the level for females was approximately 20% lower than that for males.

The average dose pattern for former smokers when they were current smokers was different from that for current smokers (Figure 9.6). At younger ages, there is evidence for an age effect, while at older ages there is evidence for a cohort effect. The peak average dose occurs at a younger age for later birth cohorts. Future periods were expected to have lower average dose for former smokers.

Estimates of cumulative consumption were calculated using consumption rates and estimates of survival for current smokers (see 3.5.4 for a derivation). These

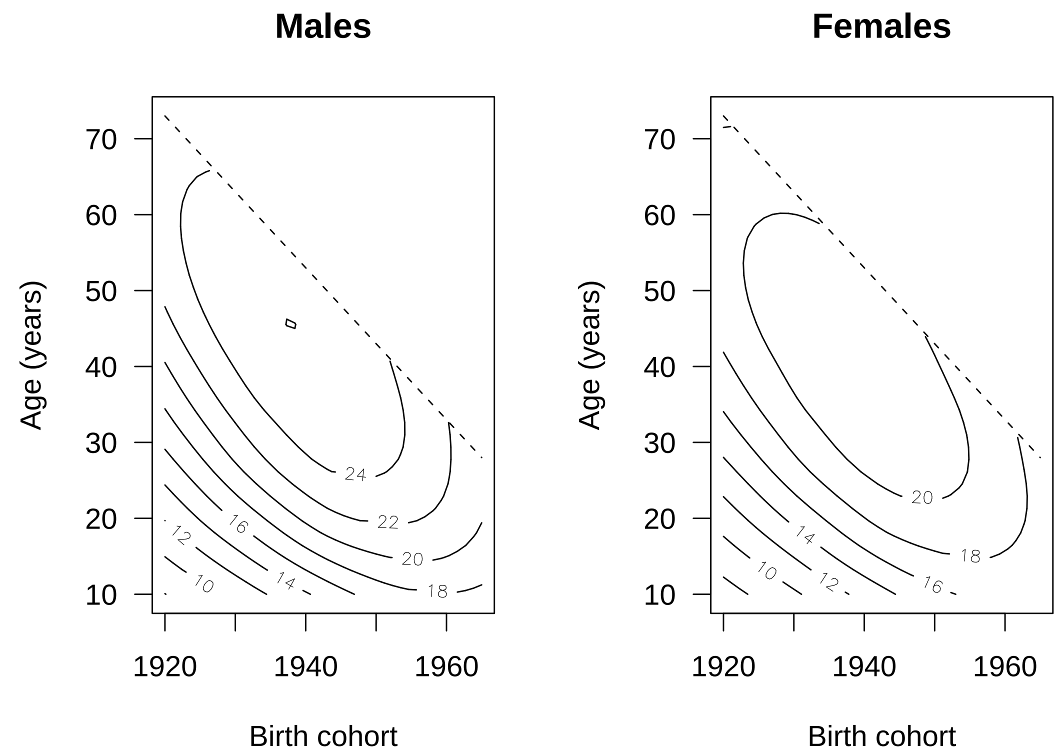


Figure 9.5: Estimated average dose (cigarettes per day) across smoking lifetime for current smokers, male and female cohorts born 1920–1970

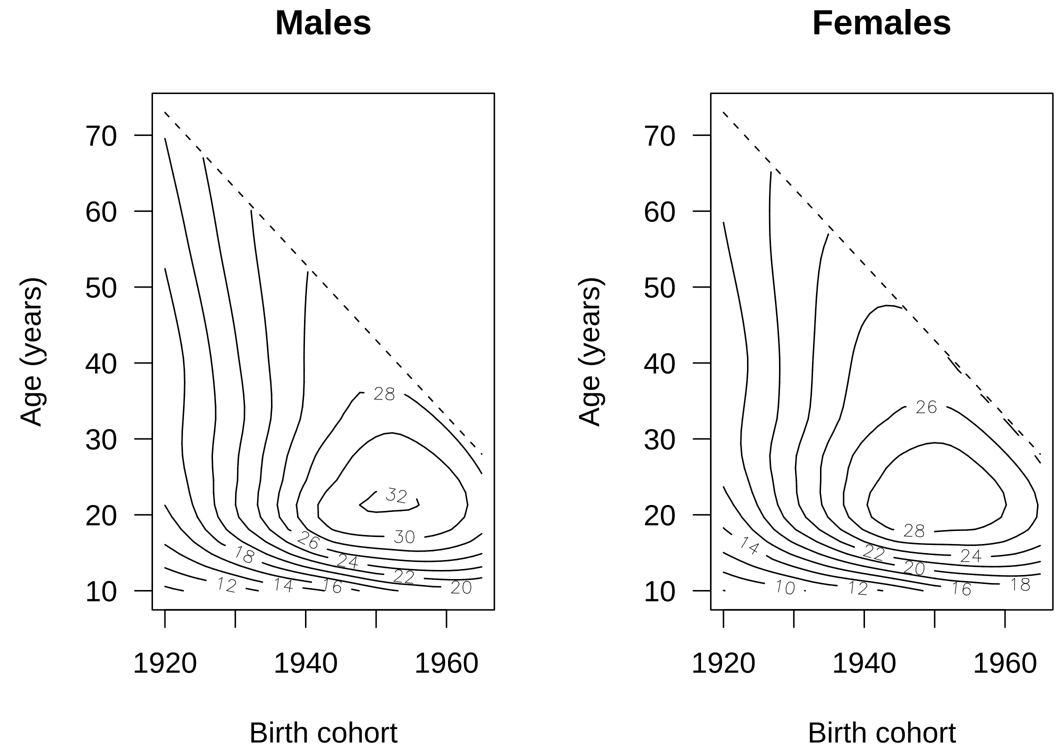


Figure 9.6: Estimated lifetime average dose (cigarettes per day) across smoking lifetime for former smokers, male and female cohorts born 1920–1970

patterns can be used to predict population-level lung cancer risk. There was a strong age effect for both males and females (Figure 9.7). Cumulative consumption for males aged 50 years peaked for the 1930 cohort, while the peak at age 30 year was for the 1940 cohort. For females, peak cumulative consumption at age 50 years was for the 1935 birth cohort, while younger ages had either peaked at the most recent period or had cumulative consumption that was rising.

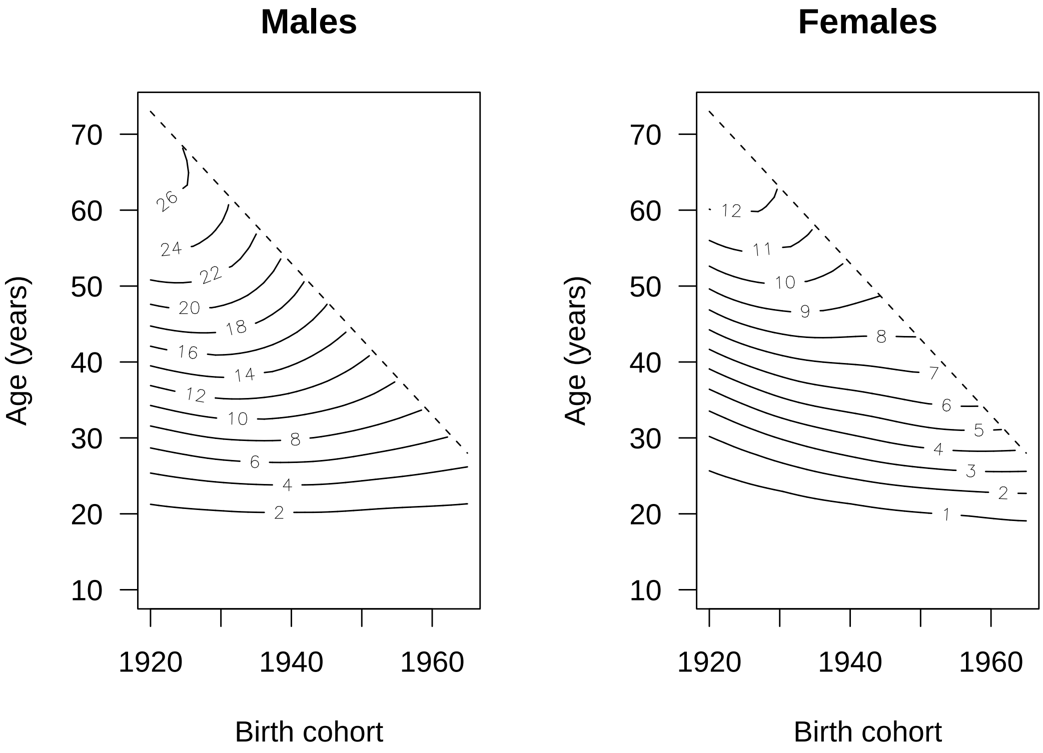


Figure 9.7: Estimated cumulative consumption (pack-years) per capita, male and female cohorts born 1920–1970

It is useful to also consider whether cumulative dose had changed for current smokers and former smokers over time and between ages. Not unexpectedly, cumulative dose estimates for current smokers increased with age (Figure 9.8). Male age-specific estimates increased over calendar periods prior to 1980, peaked, and then were stable or in slow decline from the early to mid 1980s. For females, the most recent period (dotted line is for 1995) had stable or declining age-specific rates for those aged 45 years or less, however estimates for older ages suggest rising age-specific cumulative consumption.

Cumulative consumption estimates for male former smokers followed a similar pattern to that for male current smokers for younger ages. At older ages, age-specific cumulative consumption rose for the later periods. The pattern for females was more complicated. Late uptake for the earlier birth cohorts combined with changes

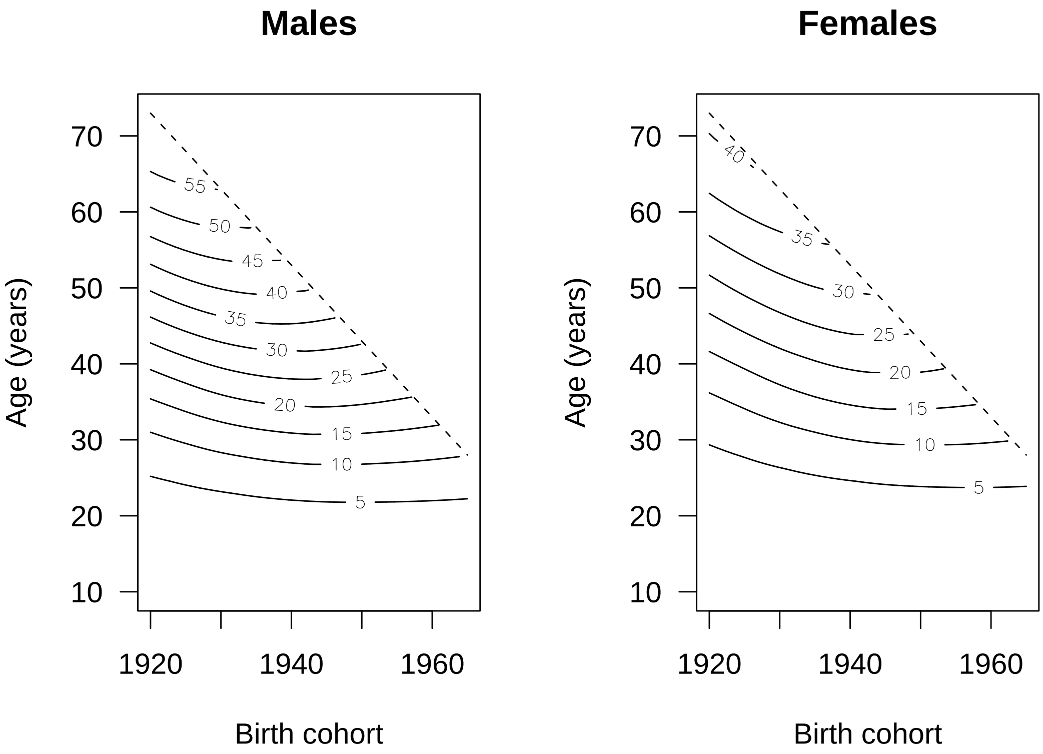


Figure 9.8: Estimated cumulative dose (pack-years) for current smokers, male and female cohorts born 1920–1970

in dose over time produced relatively stable estimates of age-specific cumulative consumption at older ages and declining estimates at younger ages. The most recent female birth cohorts also had relatively stable age-specific cumulative consumption.

9.4 Discussion

In summary, daily cigarette consumption for current smokers varied by age, between sexes and over time. Using data from 1976 to 1983, dose was relatively stable between sample surveys, with a peak in dose during middle age. Male dose tended to be higher than that for females. There was evidence to suggest that dose was highest during 1960–1975, with lower dose prior to and after that period. For the most recent periods, cumulative measures of age-specific dose were stable or dropping for both current and former smokers.

The dose estimates are potentially limited by the validity of several assumptions. First, the estimated prevalence for earlier cohorts born before 1920 was assumed to be similar to those born during 1920. Prevalence smoothing suggests that smoking prevalence for earlier cohorts may have been higher for males and lower for females.

However, external surveys for the 1950 period suggest that smoking prevalence

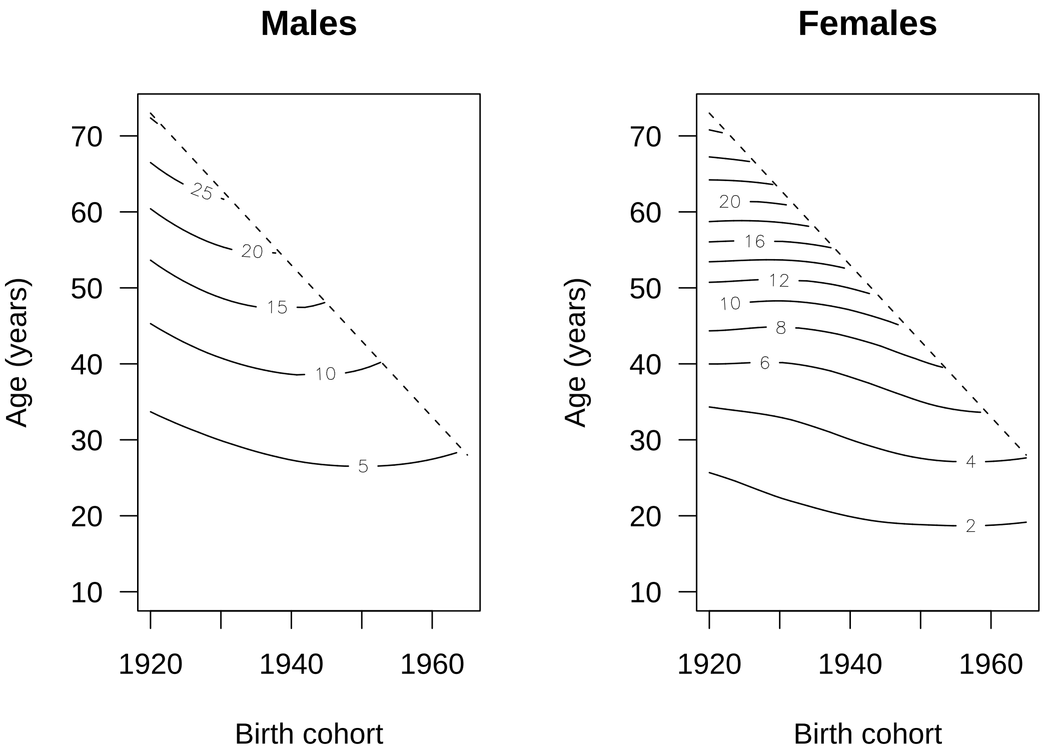


Figure 9.9: Estimated cumulative dose (pack-years) for former smokers, male and female cohorts born 1920–1970

for both males and females would have been higher, which is inconsistent with predictions for earlier cohorts. This inconsistency suggests that dose estimates prior to 1960 should be considered speculative.

A second important assumption is that age and sex specific dose is equal to dose from the 1976 New Zealand Census multiplied by a period-specific constant. It is realistic to expect that dose patterns did change over time, given historical information suggesting that females may have smoked considerably less than males in the 1930s (Tyrrell, 1999). However this assumption was made because, again, there was insufficient evidence to provide better estimates. The choice of the 1976 New Zealand Census as a reference for the variation of dose by age and sex was motivated by the absence of precise dose estimates for Australia, particularly at older ages. Moreover, the shape from the 1976 New Zealand Census data was in reasonable agreement with the Australian surveys prior to 1989.

Lower consumption rates before 1960 may be explained by several factors. First, it was more economically viable to smoke cigarettes for periods after 1960 (Tyrrell, 1999). Second, the greater use of manufactured cigarettes since the 1950s (see Figure 1.8) may lead to higher consumption per smoker: roll-your-owns are slower to smoke as they require rolling, and manufactured cigarettes have had added fire accelerants so that they are smoked more quickly (Samet, 1994). Lower consumption since the mid 1970s may be explained by smokers being more aware of the health consequences (Tyrrell, 1999).

A number of previous reports have used dose and consumption estimates for modelling lung cancer. Using methods developed by Todd (1975), British consumption data have been analysed by a number of different investigators (Townsend, 1978; Stevens and Moolgavkar, 1984; Swartz, 1992; Lee and Forey, 1998) . The consumption data were combined with smoking prevalence data available back to the 1940s and with earlier prevalence estimates from age-cohort modelling. Changes by age and by period were similar to those found in this chapter, with minor changes in the shape of age-specific dose. The British consumption data were adjusted for machine-measured tar levels, which may explain why tobacco industry-funded investigators came to the conclusion that smoking is a poor predictor of lung cancer. It is useful to note that tar-adjusted consumption declined rapidly during 1980–2000.

For Australia, Doyle (1985) estimated cumulative consumption for modelling lung cancer mortality. Consumption *per person* was assumed known up to a constant. Following assumptions about the change in male and female smoking for earlier periods, dose was estimated by scaling expected consumption compared with observed tobacco products available for consumption.

Dose has also been reported for Australia by Winstanley et al. (1995). These results are based on observed consumption and high smoker prevalence, so that estimated dose for current smokers in 1945 (3.4 cigarettes per day) was considerably lower than contemporaneous estimates in this chapter. As discussed earlier, these results are open to interpretation.

Using results from this chapter and Chapter 8, it is now possible to model population-based lung cancer mortality rates.

Chapter 10

Lung cancer mortality rate models and projections

Abstract

Aim: To fit lung cancer mortality rate models using population-based smoking exposure estimates. To project lung cancer mortality rates using projected smoking exposure estimates and the fitted regression models. **Methods:** Two regression models were fitted with terms for age, cohort and smoking exposure. The first model included data on cumulative cigarette consumption and was fitted with both generalised additive models and with generalised linear models. The second model included data on never smoker mortality rates, smoking prevalence, dose and duration. The model was fitted using generalised non-linear models and the variance was estimated using the delta method. **Results:** Lung cancer projections for ages 35–69 years from the different models were in reasonable agreement. For the middle projections of lung cancer rates, male rates were expected to continue to decline rapidly, while female rates were expected to decline more slowly. Lung cancer projections were sensitive to changes in smoking cessation, but were insensitive to changes in uptake out to 2028. **Discussion:** Lung cancer mortality rates are expected to decline and the speed of the decline is dependent on the level of smoking cessation but not on initiation rates.

10.1 Introduction

The primary motivation for the development of the multi-state smoking model was the provision of a variety of population-based smoking exposure estimates for lung cancer rate modelling (see Chapter 1). Prevalence and duration estimates were presented in Chapter 8 and dose estimates were presented in Chapter 9.

The objective of this chapter is to use the smoking exposure estimates for modelling lung cancer mortality in order to address three questions of public health significance. First, for health planning, projections are required for how lung cancer mortality is expected to change in the future. Second, for an assessment of the health impact of an intervention, projections are required for how mortality would change under different smoking scenarios, including changes for smoking initiation and for cessation. Third, there has been some recent interest in assessing whether females are at a higher risk of lung cancer mortality for a given dose and duration than are males (Prescott et al., 1998; Marang-Van de Mheen et al., 2001). It may be difficult to address this last question due to data constraints.

There have been several previous efforts to use population-based smoking exposure estimates to model lung cancer mortality rates (e.g. Townsend, 1978; Stevens and Moolgavk 1984; Doyle, 1985; Mantel et al., 1986; Tolley et al., 1991; Swartz, 1992; Holford et al., 1996; Lee and Forey, 1998; Haldorsen and Grimsrud, 1999). An important constraint on these modelling attempts has been the availability of valid and precise smoking exposure estimates.

In outline, mathematical models for carcinogenesis are reviewed and previous efforts at modelling smoking and lung cancer are described. Then several lung cancer regression models are fitted, including a comparison between lung cancer mortality rates between males and females. The fitted models are applied to different smoking exposure scenarios to estimate projected lung cancer mortality.

10.1.1 Notation and model classes

Let $\lambda(t)$ represent the mortality rate for lung cancer for age t for a given cohort [1]. The primary goal is to fit appropriate models for lung cancer mortality in terms of age, period, cohort and various smoking parameters.

Lung cancer mortality can be decomposed by smoking status. Let the lung cancer mortality rate for current and former smokers be λ_c and λ_x, respectively. Also let the lung cancer mortality rate for never smokers be represented by λ_0. Recall that the prevalence proportion for current and former smokers are π_c and π_x, respectively. Lung cancer rates for current and former smokers can be described in terms of *excess* rates compared with never smokers, such that

[1] The use of t for age and a as a constant is specific to this chapter and has been chosen for consistency with related literature. Elsewhere in this thesis, t and a have been used to represent calendar time and age, respectively.

$$\lambda = \lambda_0 + \pi_c(\lambda_c - \lambda_0) + \pi_x(\lambda_x - \lambda_0)$$

An alternative class of models is to describe the rates as a *multiplicative* model, such that

$$\lambda = \lambda_0\big[\pi_c(R_c - 1) + \pi_x(R_x - 1) + 1\big]$$

where R_c and R_x are the lung cancer mortality rate ratios for current and former smokers, respectively. The non-multiplicative models have received considerable attention because of the multi-stage theory of carcinogenesis, which is briefly reviewed in the following section.

10.1.2 Review of models of carcinogenesis

The following section reviews some models that have been used to model smoking and lung cancer. Complementary reviews are available in Forbes and Gibberd (1984) and Thomas (1988).

Multi-stage model

The multi-stage model for carcinogenesis assumes that a single cell can generate a malignant tumour only after undergoing k heritable changes (see Figure 10.1). For each person, the tissue in question is assumed to have N normal cells (at stage or state zero), each with the same independent rates (or transition intensities in the parlance of Chapter 3) for progressing through the multi-stage process. The model further assumes that the changes must occur in a specific order and are independent of age.

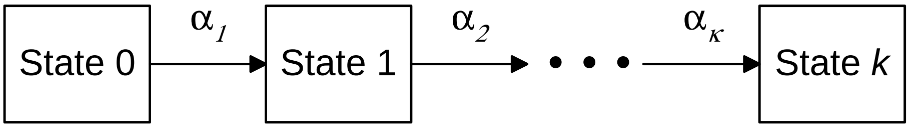


Figure 10.1: Multi-stage model for carcinogenesis with k stages

The general relationship can be found by successively conditioning upon the transitions, such that

$$\begin{aligned}\lambda(t) &= N\alpha_k(t)\int_0^t \cdots \int_0^{u_2} \\ &\quad \times P(\text{State 0 at tm 0, change to State 1 at tm } u_1, \cdots \\ &\qquad \text{change to State } k-1 \text{ at tm } u_{k-1} \,|\, \text{not in State } k \text{ by tm } t) \\ &\quad \mathrm{d}u_1 \cdots \mathrm{d}u_{k-1} \qquad (10.1)\end{aligned}$$

A reasonable simplification is to assume that the transitions are rare (Whittemore and Keller 1978), so that Equation (10.1) becomes

$$\lambda(t) \;=\; N\alpha_k(t)\int_0^t \alpha_{k-1}(u_{k-1})\cdots\int_0^{u_2}\alpha_1(u_1)\;\mathrm{d}u_1\cdots\mathrm{d}u_{k-1} \qquad (10.2)$$

For the case when the transition intensities are constant throughout life, the system is a birth process. This important case can be used to represent the lung cancer incidence rate for never smokers with no smoking exposure, such that

$$\lambda(t) = Bt^{k-1} \qquad (10.3)$$

where $B = N\alpha_1 \cdots \alpha_k/(k-1)!$

This is the familiar power relationship discussed in Armitage and Doll (1954), which can also be recognised as the incidence rate for a Weibull distribution. The probability of not reaching stage k by time t (i.e. of not becoming a malignant tumour) is

$$P(T > t) = \exp(-Bt^k)$$

This functional form subsumes a broader class of biological models (Pike, 1966).

Dependence upon exposure The transition intensities may also be dependent upon exposure status. The first case is for never smokers who become current smokers at time t_0 and smoke until time t. As will be discussed in the following section, there is good aetiological evidence that lung cancer transitions are influenced by smoking exposure at an early stage and at a late stage. Commonly, theoretical developments have assumed that the first stage and penultimate stages are affected. The transition rates are assumed constant for a given exposure status. Then the *excess* rate for current smokers is described by

$$\begin{aligned}\lambda_c(t) - \lambda_0(t) &\approx B\,r_1(t-t_0)^{k-1} \\ &\quad + B\,r_{k-1}(t^{k-1} - t_0^{k-1}) \\ &\quad + B\,r_1 r_{k-1}(t-t_0)^{k-1} \qquad (10.4)\end{aligned}$$

where the transition rates for the first and penultimate stages given exposure are scaled by r_1 and r_{k-1}, respectively (Day and Brown, 1980; Brown and Chu, 1987). The second case is for former smokers who started at t_0 and quit at time t_1, where $t_0 < t_1 \leq t$. Then

$$\begin{aligned}\lambda_x(t) - \lambda_0(t) &\approx B\,r_1\big[(t-t_0)^{k-1} - (t-t_1)^{k-1}\big] \\ &\quad + B\,r_{k-1}(t_1^{k-1} - t_0^{k-1}) \\ &\quad + B\,r_1 r_{k-1}(t_1-t_0)^{k-1}. \qquad (10.5)\end{aligned}$$

It is common to assume that r_1 and r_{k-1} have strong associations with smoking and that their values are approximately proportional to dose. Moreover, Brown and Chu (1987) estimated that the effect of smoking at the penultimate stage was approximately twice that at the first stage ($r_{k-1} \approx 2r_1$).

As the final terms in Equations (10.4) and (10.5) involve the product of r_1 and r_{k-1}, a common approximation is that

$$\begin{aligned}\lambda_c(t) &\approx \lambda_0(t) + e^{\beta}(t-t_0)^{k-1} \qquad (10.6)\\ \lambda_x(t) &\approx \lambda_0(t) + e^{\beta}(t_1-t_0)^{k-1} \qquad (10.7)\end{aligned}$$

These two equations are in similar forms, as $(t - t_0)$ and $(t_1 - t_0)$ represent smoking duration by current and former smokers, respectively.

Aetiological evidence A brief review of lung cancer aetiology was given in Chapter 1. The following section discusses lung cancer aetiology and multi-stage models. For a complementary review, see Samet (1994).

Doll and Peto (1978) fitted the multi-stage model to lung cancer incidence among British Doctors aged 40–79 years to obtain the following equation:

$$\text{incidence rate} = 0.273 \times 10^{-12}(\text{cigarettes/day} + 6)^2(\text{age} - 22.5)^{4.5} \tag{10.8}$$

An important observation by Doll (1971) was that the index of the power relationship for current smokers by duration was similar to the power relationship for never smokers by age. For Equation (10.3), the fit for current smokers from the British Doctor's Study by age gave $k \approx 7$, while by duration the fit gave $k \approx 5.5$. The latter estimate was similar to an estimate for male never smokers from CPS-I ($k \approx 5.5$).

The aetiological evidence suggests a linear relationship between the number of cigarettes smoked per day and lung cancer incidence. However the multi-stage model predicts a quadratic dose relationship because there is evidence for both an early effect, as suggested by duration, and for a late stage effect, due to reduced risk following smoking cessation. This issue was raised by Armitage (1971) in response to a paper by Doll (1971). Doll and Peto (1978) sought to address this issue in terms of both random error and possible systematic biases (see also Lee, 1995).

Evidence suggests that the absolute risk of lung cancer following cessation remains relatively constant. One of the difficulties in estimating the effect of cessation is that smokers may quit for different reasons, including ill health. Moolgavkar et al. (1989) suggest "In fact, interpretation of the data on risk among exsmokers is quite controversial, and the only certainty is that this risk lies somewhere in between the risk among continuing smokers and that among nonsmokers."

Two-mutation model

The multi-stage model has received considerable attention, possibly due to its generality. However some investigators have criticised the multi-stage model as providing inconsistent models when fitted to current smokers and to former smokers separately (Gaffney and Altshuler, 1988). An alternative model is the two-mutation "recessive oncogenesis" model.

Let $X(t)$ be the mean number of susceptible stem cells, which increases with age t. Also let the number of intermediate cells be Y and the number of malignant cells be Z, which can be expressed in terms of X and the rates of change. The variable $d(t)$ is the smoking dose, which is a function of age t. For never smokers, $d(t) = 0$ for all t. For current smokers, $d(t) = 0$ for $t < t_0$ and $d(t) = d$ (constant) for $t \geq t_0$.

The mutation rates from susceptible cells to intermediate cells (including splitting) is $c_0 + c_1 d(t)$ and from intermediate to malignant cells is $c_0 + c_2 d(t)$. Finally, the net proliferation rate for intermediate cells is $a + bd(t)$.

The model is shown graphically in Figure 10.2.

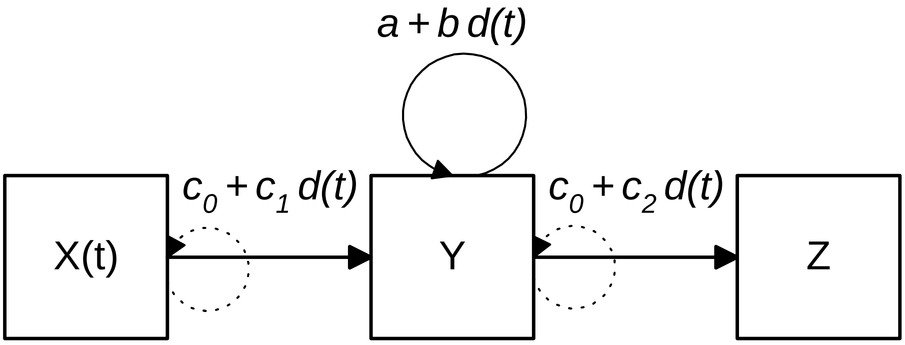


Figure 10.2: Two mutation model for carcinogenesis dependent upon changing dose

The hazard of interest is the product of the number of intermediate cells times the transition rate from intermediate to malignant cells, conditional upon no earlier malignant cells ($[c_0 + c_2 d(t)]\mathrm{E}(Y(t)|Z(t) = 0)$). By not conditioning upon $Z(t) = 0$, the hazard can be expressed by:

$$\begin{aligned} h(t,d) &= (c_0 + c_2 d)\Bigg\{ c_0 \exp\big[bd(t-t_0)\big] \int_0^{t_0} X(s) \exp\big[a(t-s)\big] \mathrm{d}s \\ &\quad + (c_0 + c_1 d) \int_{t_0}^{t} X(s) \exp\big[(a+bd)(t-s)\big] \mathrm{d}s \Bigg\} \qquad (10.9) \end{aligned}$$

where $h(t, d)$ is the lung cancer hazard for age t and d cigarettes per day (Moolgavkar et al., 1989, Equation (1)).

For model simplifications, assume that $X(t) = X(t_0)$ (constant) for $t \geq t_0$, which closely follows the development in Moolgavkar et al. (1989), who had that $t_0 = 19$ years and $X(t)$ was constant from age 20 years. If we also assume that $X(t) = X(t_0/2)$ (constant) for $t < t_0$, then Equation (10.9) simplifies to

$$\begin{aligned} h(t,d) &\approx (c_0 + c_2 d)\Bigg\{ c_0 X(t_0/2) \exp\big[bd(t-t_0)\big] \frac{\exp(at) - \exp\big[a(t-t_0)\big]}{a} \\ &\quad + (c_0 + c_1 d) X(t_0) \frac{\exp\big[(a+bd)(t-t_0)\big] - 1}{a+bd} \Bigg\} \end{aligned}$$

This suggests that the hazard for never smokers will be dominated by an exponential for age, while the hazard for current smokers is composed of both a product between a linear dose term and exponential age together with a product between a quadratic term for dose and an exponential for duration.

Moolgavkar et al. (1989) fitted a two-mutation model for never smokers compared with current smokers. They found that different models fitted the British Doctors Study well, some of which suggest that age may have an independent influence on lung cancer risk. They also found that the relationship between dose and duration was highly model-dependent. The authors also criticised the simplified model proposed by Gaffney and Altshuler (1988) and recommended that models for former smokers be re-fitted when reliable data became available. In summary, issues related to modelling lung cancer mortality for formers smokers were not resolved.

One-stage models

In the presence of equivocal evidence, an alternative approach is to use simpler models (Forbes and Gibberd, 1984). One such model assumes that carcinogenesis follows only after reaching a critical value for the number of mutated cells (Brown and Forbes, 1974). By further assuming that the number of mutated cells is normally distributed and that the risk of cell mutation is a linear function of both period and cumulative consumption, the lung cancer incidence rate can be represented by

$$\lambda(t) = 1 - \Phi(\beta_0 + \gamma t + \theta \text{CumCons}(t))$$

where $\Phi()$ is the cumulative normal function, β_0, γ, θ are constants, and $\text{CumCons}(t)$ is cumulative consumption at age t. Although this model is simple and based on a sound biological model, the model has received little attention.

10.1.3 Population-based lung cancer models

The most valid study design for modelling changes in smoking and lung cancer mortality is to use a large aetiological study design such as a cohort or case-control study. However the costs for such studies are often prohibitive. One alternative is to use lung cancer mortality rates from vital statistics together with population-based smoking exposure information.

This aggregate study design has the advantage of low cost and the potential for high precision, as the number of lung cancer cases can be large. However the approach has several limitations. First, there is potential for an *ecological fallacy*, where an aggregate change may not be observed evenly across a stratum. As a hypothetical example, a smoking intervention may not reach lower socio-economic groups, who may be at increased occupational or domestic exposure to lung cancer risk. Any decline in smoking exposure in the general population may not then translate to a commensurate change in lung cancer incidence. However in the absence of good evidence for differential distribution of other risk factors, fallacious ecological associations have been assumed negligible. In particular, it is assumed that a population-level intervention will be homogeneous across a population.

Second, population-based measures of smoking exposure may be inaccurate, limiting the *precision* and *validity* of the lung cancer models. Precise and valid survey data are required together with appropriate estimates of smoking to measure population exposure heterogeneity. However the nature of tobacco exposure has changed rapidly over time. This may explain, in part, why tobacco industry funded researchers used population-based data to conclude that smoking was a poor predictor of lung cancer mortality (Swartz, 1992; Lee and Forey, 1998).

Third, the population-based aggregate measures of smoking exposure are limited by their ability to represent the *heterogeneity* of risk of lung cancer in a population.

As a consequence, the formulations for mathematical models of carcinogenesis described in the previous section tend to be simplified for the purposes of population-based lung cancer modelling. Moreover, the observation from individual-based studies that different models may fit a set of data equally well (see Moolgavkar et al., 1989; Thomas, 1988) holds equally true for population-based models.

In the following section, some previous efforts at population-based lung cancer models are briefly described.

Multi-stage models

There have been several applications of the multi-stage model to population-based data.

Townsend (1978) extended the multi-stage model to include effects for cigars and pipes. The approach can be criticised for making broad data assumptions about the historical smoking exposure, where the patterns of age-specific and sex-specific prevalence for 1948–1960 were assumed to apply back to 1886. One advantage of the model was the incorporation of the full distributional form for duration of current smoking and time since cessation, rather than relying on the early order moments.

The first smoking monograph from the National Cancer Institute included a sophisticated lung cancer model (Tolley et al., 1991). The authors used a multi-state smoking model, with compartments including age and smoking duration. The smoking estimates were used to fit a simple multi-stage lung cancer rate model following forms suggested by Peto (1986). The multi-state model was projected out in time, taking account of smoking initiation and cessation, lung cancer mortality and competing causes of death. This approach has a number of merits, including explicit incorporation of smoking duration and competing risks, however it has not been widely used.

Industry-funded researchers have used the multi-stage model with cumulative estimates of British tobacco consumption adjusted for tar (Swartz, 1992; Lee and Forey, 1998). Their research is notable for the careful model development and use of data. Forey et al. (1998) comment as to the data limitations of the constant cumulative tar estimates, including possible bias due to the measurement of tar levels.

Holford et al. (1996) presented a careful mathematical development for population-based lung cancer modelling. Their development for population smoking exposure heterogeneity is described in Section 10.1.3. Both additive and multiplicative models were fitted, although the authors found the multiplicative models had better convergence properties. The fitted parameters were inconsistent, possibly because the models were ill-conditioned.

A recent analysis by Haldorsen and Grimsrud (1999) used Norwegian data to fit a multi-stage model. The model included excess rate parameters for dose and duration, with some account for the time since cessation. Smoking data for Norway were available from different smoking surveys going back to the 1950s. The study provides a good example of population-based lung cancer mortality modelling.

Two-mutation models

Although the two-mutation model has been recommended for investigating certain hypotheses for carcinogenesis (Thomas, 1988), the models require good information on dose and duration. This has limited the application of these models to population-based lung cancer modelling.

One-stage models

The probit model from Section 10.1.2 has been used to model lung cancer mortality in Canada (Mantel et al., 1986) and Australia (Doyle, 1985). These applications have tended to emphasise the modelling for a given birth cohort, without modelling across the Lexis diagram. Across the Lexis diagram for Australian data, the probit model did not perform as well as a Poisson regression model with log link and with a linear term for cumulative consumption (see the Commonwealth report in Appendix A).

Stevens and Moolgavkar (1984) developed a statistical model based on cumulative cigarette consumption and ever smoker prevalence. Let $R(x)$ be the lung cancer mortality rate ratio due to cumulative dose of x units. Then the lung cancer rate for current smokers is

$$\lambda_c(t) = \lambda_0(t) \int_0^\infty R(x) \mathrm{d}P_{\mathrm{CumDose}}(x).$$

where P_{CumDose} is the distribution for cumulative dose. The authors assumed that risk for a given level of exposure was *multiplicative*, where $R(x) = \beta^x$. Then, given moderately small second order and higher moments, the lung cancer mortality rate is approximated by

$$\lambda_c(t) \approx \lambda_0(t)\beta^{\overline{\mathrm{CumDose}}}.$$

For the purposes of population-based modelling of lung cancer mortality, this suggests a linear term for a log-link, which is at variance with the multi-stage model that suggests a log term for cumulative consumption for a log-link.

Model selection

Selection of a model for lung cancer mortality and smoking is constrained by the availability and precision of the exposure data. For a consumption-based model, the one-stage model due to Stevens and Moolgavkar (1984) has the advantage of providing estimates for population attributable fractions, however estimates of ever smoking prevalence were also required. Poisson regression models provide a flexible modelling framework for cumulative consumption (see Appendix A). A useful choice for the functional form for cumulative consumption would be to follow that proposed by Stevens and Moolgavkar (1984) for the total population.

The multi-stage model and two-mutation model both require data on prevalence, duration and potentially dose. The multi-stage model has often been used for population-based modelling, possibly due to its simplicity.

Time from malignancy to death

Cell transformations to cancer will have taken place several years before a lung cancer death. One consequence is that the relevant exposure is exposure up to the time of the pre-clinical onset of cancer. This relationship is represented in Equation (10.8) by the expression (age − 22.5) where the mean age of uptake was 19.2 years and only a 3.3 years were expected before the cancer was clinically evident (Doll and Peto, 1978).

Lung cancer is an aggressive disease, where the time between clinical onset and death is small. Whittemore (1988) used 5 years as the time between carcinogenic transformation and death. Following this approach, the age effect was parameterised for (age$-$5) and estimates of dose and duration for Model B were only considered up to the five years prior to observation. The adjusted estimates for dose and duration are represented by Dose* and Dur*, respectively. For model simplicity, cumulative consumption was not lagged.

Population heterogeneity in smoking exposure

From Equations (10.6)–(10.7) together with a linear dose term, the lung cancer mortality rate is conditional upon the adjusted values of dose and duration, such that

$$\begin{aligned}\lambda_c(t) &= \lambda_0(t) + e^{\beta}\text{Dose}^*_c(\text{Dur}^*_c)^{\theta}\\ \lambda_x(t) &= \lambda_0(t) + e^{\beta}\text{Dose}^*_x(\text{Dur}^*_x)^{\theta}\end{aligned}$$

at age t for current and former smokers, respectively. The linear dose term was chosen in place of a quadratic term for model simplicity. Doll and Peto (1978) found that the choice was equivocal, while Haldorsen and Grimsrud (1999) estimated the power as 1.55, which is equivocal. The linear term simplifies any account for population heterogeneity.

The population-based lung cancer mortality rate is the expected value over the dose and duration distributions. Following the development in Holford et al. (1996), the expected value can be approximated using a Taylor's series expansion about the first and second moments. By assuming no correlation between dose and duration, the expansion gives

$$\lambda_c(t) = \lambda_0(t) + e^{\beta}\text{E}(\text{Dose}^*_c)\cdot\text{E}(\text{Dur}^*_c)^{\theta}\left[1+\frac{1}{2}\theta(\theta-1)\frac{\text{var}(\text{Dur}^*_c)}{\text{E}(\text{Dur}^*_c)^2}\right]$$

with a similar expansion for former smokers. The term $\text{var}(\text{Dur}^*_c)/\text{E}(\text{Dur}^*_c)^2$ is the square of the coefficient of variation.

10.2 Methods

10.2.1 Data sources

Population estimates and the number of lung cancer deaths by five year age groups and single calendar year were available for 1950–1999. See Chapter 2 for further details.

Estimates of exposure for cohorts born 1915–1965 for periods until 1993 were available from Chapters 8 and 9. These data were aggregated into five-year age groups by weighting the estimates by the population size for single years of age. Estimates for cohorts born 1915–1919 were assumed similar to the 1920 birth cohorts. Estimates for 1994–1999 calendar period were assumed to have similar transition rates to 1993.

The Lexis diagram for the data used in the lung cancer mortality regression is shown in Figure 10.3.

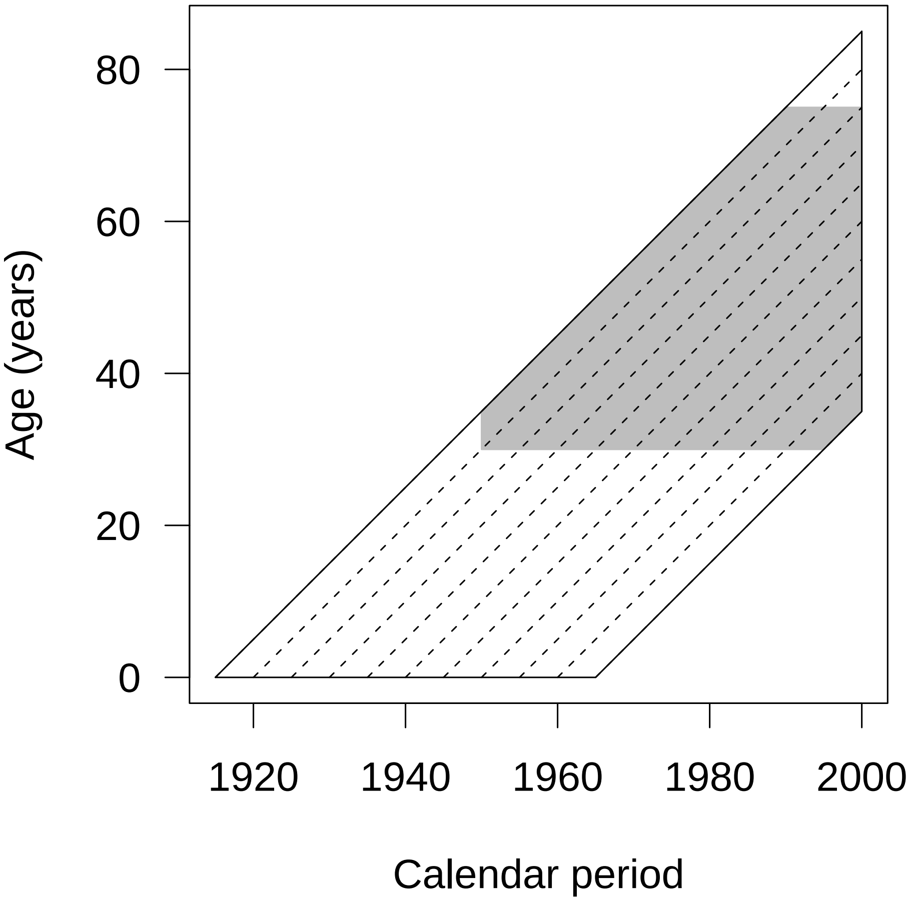


Figure 10.3: Lexis diagram of the cohorts used for model fitting (grey shading=lung cancer mortality rates used)

10.2.2 Lung cancer regression models

Two groups of models were used for the lung cancer regression. The first group used estimates of population cumulative cigarette consumption, while the second group used prevalence, dose and duration estimates for current and former smokers. Both models included parameters for age, cohort and tobacco exposure. Importantly, period effects were not estimated because the retrospective study design constrained the available cohorts across the Lexis diagram so that any period effect would be confounded with *reversed* age and cohort effects.

The first regression model included effects for age, cohort and cumulative cigarette consumption (CumCons) per *person*:

Lung cancer Model A

$$\lambda \;=\; \exp\left[\beta_0 + \alpha(\text{age}) + \nu\,(\text{cohort}) + \theta(\text{CumCons})\right]$$

This model was based on the one-stage models for a population, without requiring prevalence data. The model was fitted as a generalised additive model with the covariate functions $\alpha()$, $\nu()$ and $\theta()$ represented by spline functions.

Inspection of the forms of these functions allowed the models to be fitted as generalised linear models. Two forms were considered for age. First, age was modelled as a factor, with a set of indicator variables. Second, age was modelled using a log function so that

$$\exp\left[\alpha(\text{age})\right] = (\text{age} - 5)^{\alpha}$$

where the adjustment for five years is for the approximate time for transformation to cancer (see Section 10.1.3).

Model A summarises smoking exposure by the cumulative consumption *per caput* of population. The other lung cancer model decomposes smoking into current and former smokers, including a linear term for dose and estimates the number of stages for duration (Model B).

Lung cancer Model B

$$\begin{aligned}\lambda &= \exp\left[B_0 + \nu(\text{cohort})\right] \\ &\quad\times\Bigg\{(\text{age}-5)^A + \\ &\quad\pi_c \exp(\beta)\text{E}(\text{Dose}^*_c)\text{E}(\text{Dur}^*_c)^\theta\left[1+\frac{1}{2}\theta(\theta-1)\frac{\text{var}(\text{Dur}^*_c)}{\text{E}(\text{Dur}^*_c)^2}\right] + \\ &\quad\pi_x \exp(\beta)\text{E}(\text{Dose}^*_x)\text{E}(\text{Dur}^*_x)^\theta\left[1+\frac{1}{2}\theta(\theta-1)\frac{\text{var}(\text{Dur}^*_x)}{\text{E}(\text{Dur}^*_x)^2}\right]\Bigg\}\end{aligned}$$

This model is based on the multi-stage model with an allowance for population heterogeneity. The covariate ν(cohort) was represented as a polynomial with order as suggested by the generalised additive model analysis of Model A. The values for B_0 and A were taken from CPS-I (see Section 10.2.3).

10.2.3 Estimates for never smokers

Fitting the population-based lung cancer risk equations has been recognised as being difficult (Holford et al., 1996). One possible solution is to use exogenous lung cancer rates for never smokers, which is an approach that has been used by Townsend (1978) and by Whittemore (1988).

Data were taken from the twelve year follow-up from CPS-I (Burns et al., 1997c). For comparison, data were also taken from the six-year follow-up from CPS-II (Thun and Heath, 1997). The data were approximately mid-period for the lung cancer rates from the cohorts born after 1920 for the period 1950–1994. Following the approach by Whittemore (1988), regression models were fitted for white males and females.

The model chosen was

$$\begin{aligned}\lambda_0(\text{age}) &= \exp\left[\mu + \alpha\log(\text{age}-5)\right] \\ &= e^\mu(\text{age}-5)^\alpha\end{aligned}$$

This can be fitted as a generalised linear model with Poisson error structure, log link and an offset for the log of the population (Breslow and Day, 1987).

The fitted equations are shown in Table 10.1. The smoothed curves are shown in Figure 10.4. In a comparison of fitted values between CPS-I and CPS-II, the

	Study	
Group	CPS-I	CPS-II
Males	$0.201 \times 10^{-12}(\text{age} - 5)^{5.09}$	$2.54 \times 10^{-15}(\text{age} - 5)^{6.08}$
Females	$0.981 \times 10^{-12}(\text{age} - 5)^{4.60}$	$0.670 \times 10^{-12}(\text{age} - 5)^{4.70}$
Total	$0.672 \times 10^{-12}(\text{age} - 5)^{4.71}$	$0.180 \times 10^{-12}(\text{age} - 5)^{5.02}$

Table 10.1: Never smoker lung cancer mortality rate equations based on CPS-I and CPS-II data, by sex

rates for older males and for females of all ages were in reasonable agreement. The comparison for males suggested a steeper slope for CPS-II at older ages.

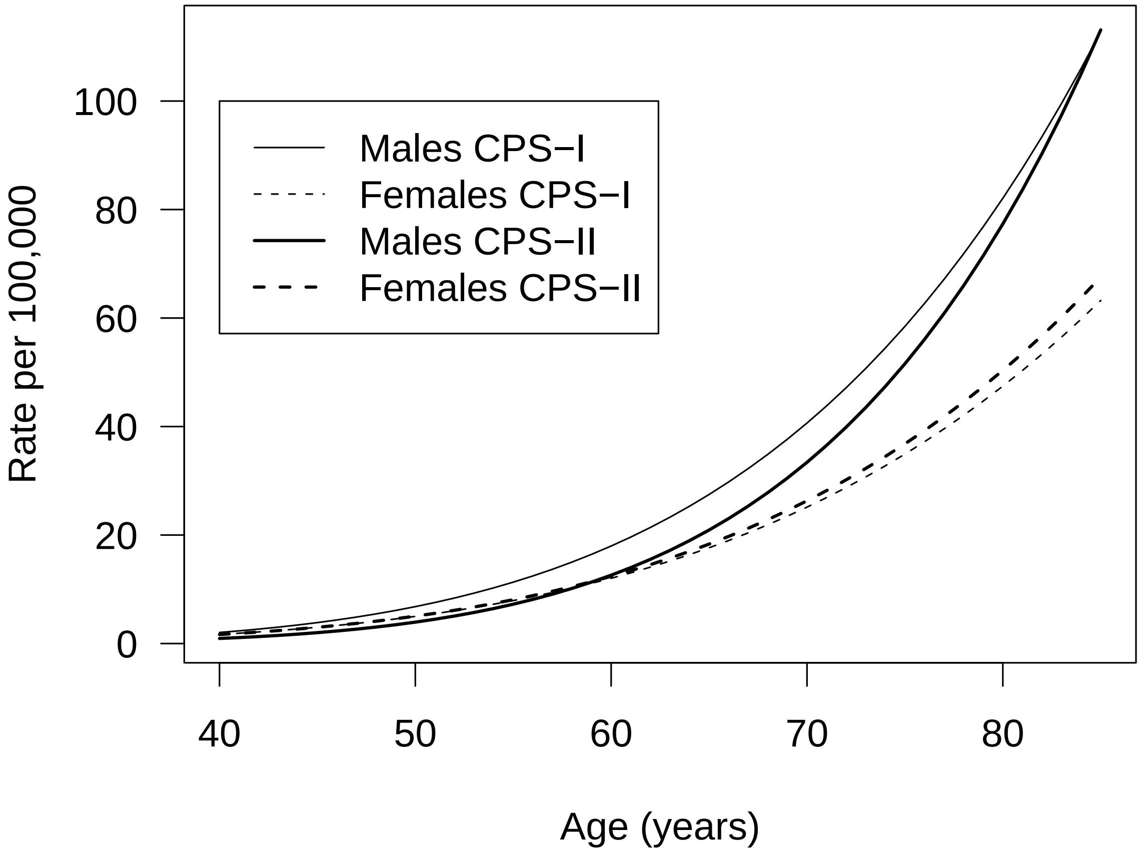


Figure 10.4: Fitted never smoker lung cancer mortality rates, CPS-I and CPS-II

10.2.4 Model fitting

For all models, the numbers of deaths were assumed to follow a Poisson distribution (Brillinger, 1986).

Model A was fitted using generalised additive models and generalised linear models with a log link. The generalised additive models were fitted using the `gam()`

function in the mgcv package in R, which estimated the smoothing parameters using generalised cross validation (Wood, 2000). The generalised linear models used the glm() function in R. Moreover, the generalised linear models included the cohort function $\nu()$ as a polynomial in terms of cohort-1950 for numerical stability.

Model B was fitted using non-linear maximum likelihood estimation. Computations used the gnlr() function in the R package gnlm ("generalised non-linear models"). The package allowed for flexible model formulation given a Poisson error distribution. The gnlr() function used the nlm() maximisation function which is based on a Newton-like algorithm. Initial values for the non-linear estimation were taken from more simple models, where available. Initial values were taken from a range of values to find stable estimates.

Goodness of fit statistics were reported using deviance statistics and Akaike's Information Criterion, together with degrees of freedom.

10.2.5 Variance estimation

Variance estimation for predictions varied depending upon the model. By recognising that Model A is a generalised linear model, standard methods were used for estimation of the variance of the linear predictor (see McCullagh and Nelder, 1989). Confidence intervals for the mean number of deaths were found by

$$\exp(\hat{\beta}'x \pm z_{(1-\alpha/2)}\sqrt{x'\Sigma_{\hat{\beta}}x})$$

with variance estimated using the approximation

$$\exp(\hat{\beta}'x)^2 \ (x'\Sigma_{\hat{\beta}}x).$$

Rates were estimated using the mean number of deaths divided by the population. The variance for the predicted mean from Model B was estimated using the delta-method with validation using the bootstrap, as discussed in the following two sections.

Delta method

Estimates of the covariance matrix for the predicted values for Model B can be estimated using the delta method (Rao, 1973). Specifically, let the set of parameters be

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \pi_c \\ \pi_x \\ \text{Dose}_c^* \\ \text{Dose}_x^* \\ \text{Dur}_c^* \\ \text{Dur}_x^* \end{pmatrix}$$

where Dose_c^* and Dur_c^* are the 5-year lagged estimates for dose and duration among current smokers, respectively, and Dose_x^* and Dur_x^* are the comparable estimates for former smokers. Let the point estimate as a function of the parameters be

$$g(\boldsymbol{\theta}) = \log[\lambda(\boldsymbol{\theta})].$$

For notational simplicity, let

$$\begin{aligned} \boldsymbol{\Sigma}_\beta &= \text{var}(\boldsymbol{\beta}) \\ \boldsymbol{\Sigma}_\pi &= \text{var}\begin{pmatrix} \pi_c \\ \pi_x \end{pmatrix} \\ \boldsymbol{\Sigma}_{D^*} &= \text{var}\begin{pmatrix} \text{Dose}_c^* \\ \text{Dose}_x^* \\ \text{Dur}_c^* \\ \text{Dur}_x^* \end{pmatrix} \end{aligned}$$

so that the covariance matrix for the parameter vector $\boldsymbol{\theta}$ is

$$\boldsymbol{\Sigma} = \text{var}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Sigma}_\beta & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_\pi & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_{D^*} \end{pmatrix}.$$

By introducing the matrix of partial derivatives

$$\boldsymbol{G} = \left(\frac{\partial g_i}{\partial \theta_j}\right),$$

we can estimate the desired variance

$$\text{var}(g) = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}'.$$

In practice, the partial derivatives were calculated numerically using `R`. For comparison with the bootstrap results, the variances for prevalence and the exposure measures were assumed negligible. The delta method was used as the primary estimates for variance estimation. The estimates were validated using the bootstrap.

Bootstrap

It is useful to note that, as I am working with generalised non-linear models, developments for bootstraps of generalised linear models apply. The main difficulty with the bootstrap is dealing with the heteroscedasticity, requiring that the regression residuals be scaled (Moulton and Zeger, 1991; Davison and Hinkley, 1997).

First define $\hat{d}_i = \hat{\lambda}_i \, \mathrm{Pop}_i$, which is the expected number of deaths. Friedl (1997) suggested that the scaled residuals e_i use a modified form of the Pearson's residuals which has similar asymptotic properties to deviance residuals, such that

$$e_i = \frac{d_i - \hat{d}_i}{\sqrt{\hat{d}_i(1 - p/n)}}$$

where p is the number of parameters and n is the number of observations. The residuals were centred ($e_i \leftarrow e_i - \bar{e}$). By resampling e_i^* from $\{e_i\}$, the resampled deaths are

$$d_i^* = \hat{d}_i + e_i^* \sqrt{\hat{d}_i}.$$

Model B is then re-fitted for the resampled deaths, and the means for the predictions used to estimate standard errors. The mechanical details of the bootstrap were dealt with using the `boot` package in `R`.

For validation of the standard errors, 100 replications were used. When the models for the re-sampled data did not converge or produced outlying parameter estimates more than four standard deviations from the mean, then alternative starting points were considered. Confidence intervals using standard errors based on the delta method were not graphically different from confidence intervals using standard errors based on the bootstrap. A useful extension would be to use the 2.5% and 97.5% confidence intervals, however the data were reported using age-standardised rates, which required variance estimates.

Variance estimates for age-standardised rates

Following Hakulinen and Dyba (1994), variance estimates for age-standardised rates based on rate models require the incorporation of covariance estimates. Let $\Sigma_{\hat{\lambda}}$ be the covariance matrix for mean predicted rate, with standard population weights w, and populations n. Then the variance for the mean age-standardised rate was calculated as

$$w'\Sigma_{\hat{\lambda}}w$$

and the variance for individual predictions was estimated using

$$w'\Sigma_{\hat{\lambda}}w + (w^2)'\frac{\hat{\lambda}}{n}.$$

Calculations were performed for the generalised linear models and the generalised non-linear models. Covariance estimates were not available for the generalised additive models.

Comparisons between males and females

Model B was fitted for several models to test for differences between males and females. First, Model B was extended to include a multiplicative indicator variable for females. The fitted parameter was then interpreted as a rate ratio for females compared with males. Second, Model B includes terms for the constant term for smokers (β) and the power term for duration (θ) being different for females compared with males. Third, a model was fitted including all three of these terms.

10.2.6 Projections

Projected smoking exposure

Smoking exposure was projected under four different scenarios:

1. Middle projection (transition rates as per 1993)
2. Increase in cessation ($\alpha_Q^* = 2\alpha_Q$)
3. Decrease in cessation ($\alpha_Q^* = \alpha_Q/2$)
4. No uptake ($\alpha_I^* = 0$)

where α_I^* and α_Q^* are the revised rate estimates for smoking initiation and cessation, respectively. These rates were entered into the smoking model and then the smoking model was pushed forward in time. See Chapters 8 and 9 for details. The projections were performed for the 1915–1995 birth cohorts (see the Lexis diagram in Figure 10.5). The transition rates for birth cohorts born after 1965 for periods before the 1993 calendar year were smoothed.

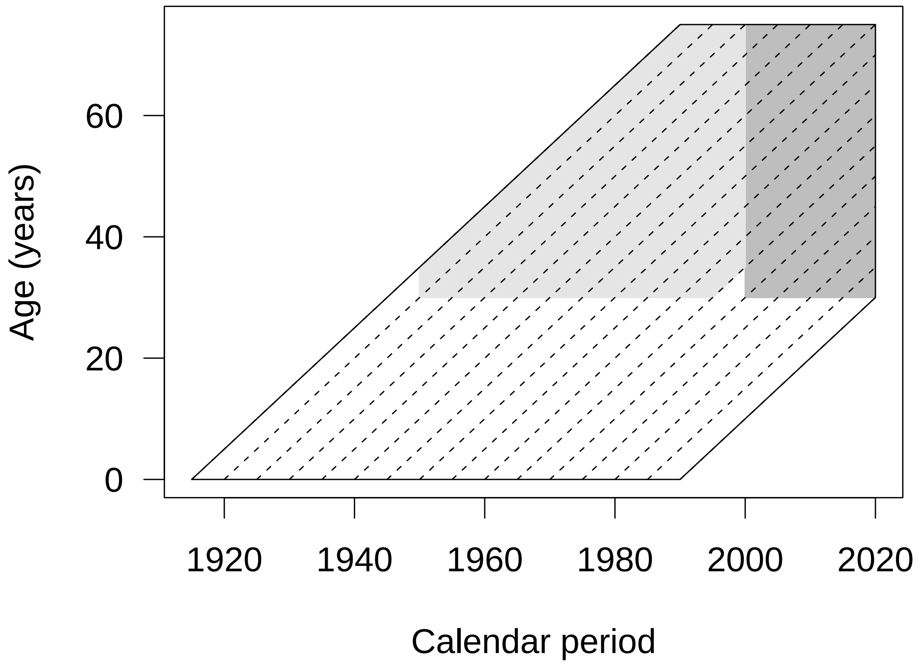


Figure 10.5: Lexis diagram of the cohorts for which model projections were made (darker grey shading = projected lung cancer mortality rates; lighter grey shading = observed lung cancer mortality rates)

Projected lung cancer mortality

Middle lung cancer projections used the middle smoking projection applied to Model A, fitted for both generalised linear models and generalised additive models, and Model B. An age-period generalised additive model using a two-dimensional spline was also presented for cross-method validation. The other lung cancer scenarios were based on the associated projected smoking exposure applied to Model B.

10.3 Results

Generalised additive models (GAMs) were used to explore the functional form for Model A (see Figures 10.6 and 10.7). The effects shown in the figures assume a log-link.

For the age effect, both males and females show evidence for curvature, suggesting that log(age − 5) would be a suitable functional form. Moreover, the effect for cumulative consumption is linear for males and linear or convex (concave upwards) for females, which supports using a linear effect rather than a log effect as suggested by the multi-stage model.

The interpretation of the cohort effects is more complicated. The cohort effects are following an adjustment for age and cumulative consumption, representing unexplained variation. The male cohort effects are relatively stable from earliest birth cohorts to the 1950 birth cohort and then declines for the later birth cohorts. The female cohort effects rise for cohorts born from 1920 to 1950 and then drop in a manner similar to males. The small drop from the 1950 cohorts may be explained by lack of formal adjustment for tar. The earlier rise for females is difficult to interpret and may be due to a bias in the exposure estimates for females. The effects for both males and females suggest curvature, possibly with a quadratic functional form.

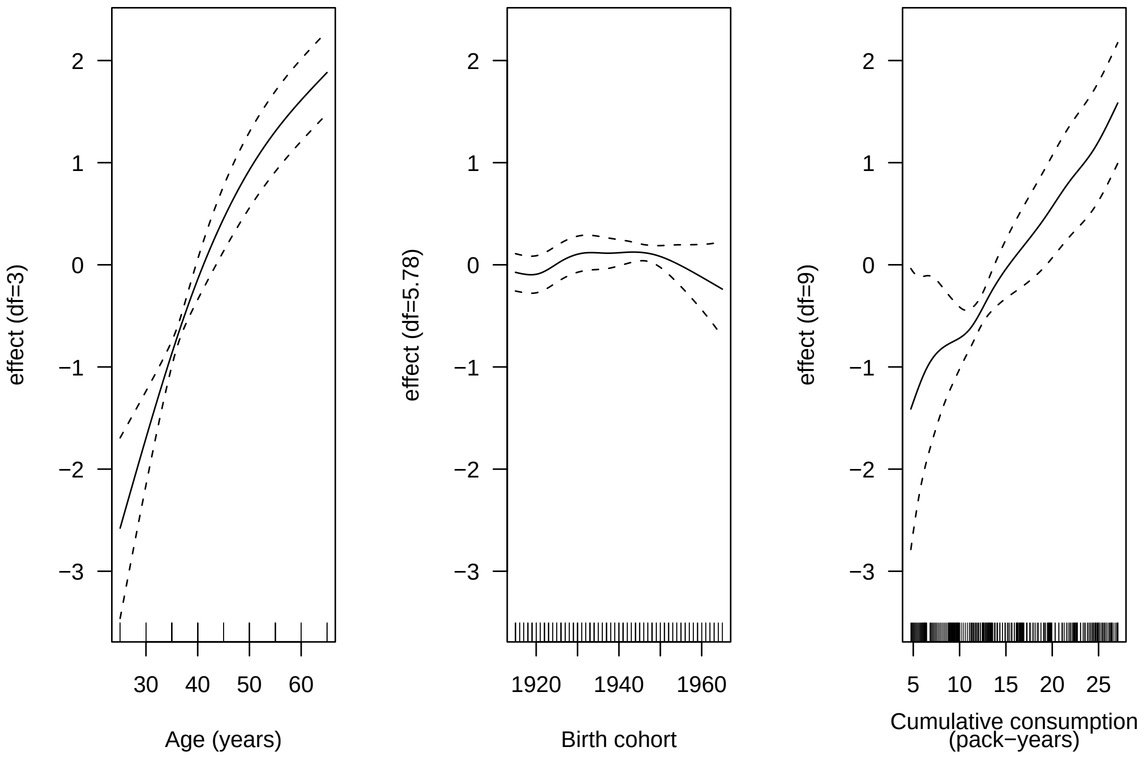


Figure 10.6: Smoothed covariate parameters for the GAM fit for Model A, males

Given the implied functional forms from the GAMs, generalised linear models (GLMs) were fitted for formal model comparisons. The model comparisons are summarised in Tables 10.2–10.4. The models for non-parametric age include age as a factor, while the models with parametric age include the function log(age-5) to model age. For each model, the degrees of freedom (df), the residual deviance Akaike's Information Criterion (AIC) are presented. The cohort terms are included either as linear or quadratic polynomial (=quadratic(cohort)) terms. For Model A, the smoking parameter is a linear term for cumulative consumption. For Model B,

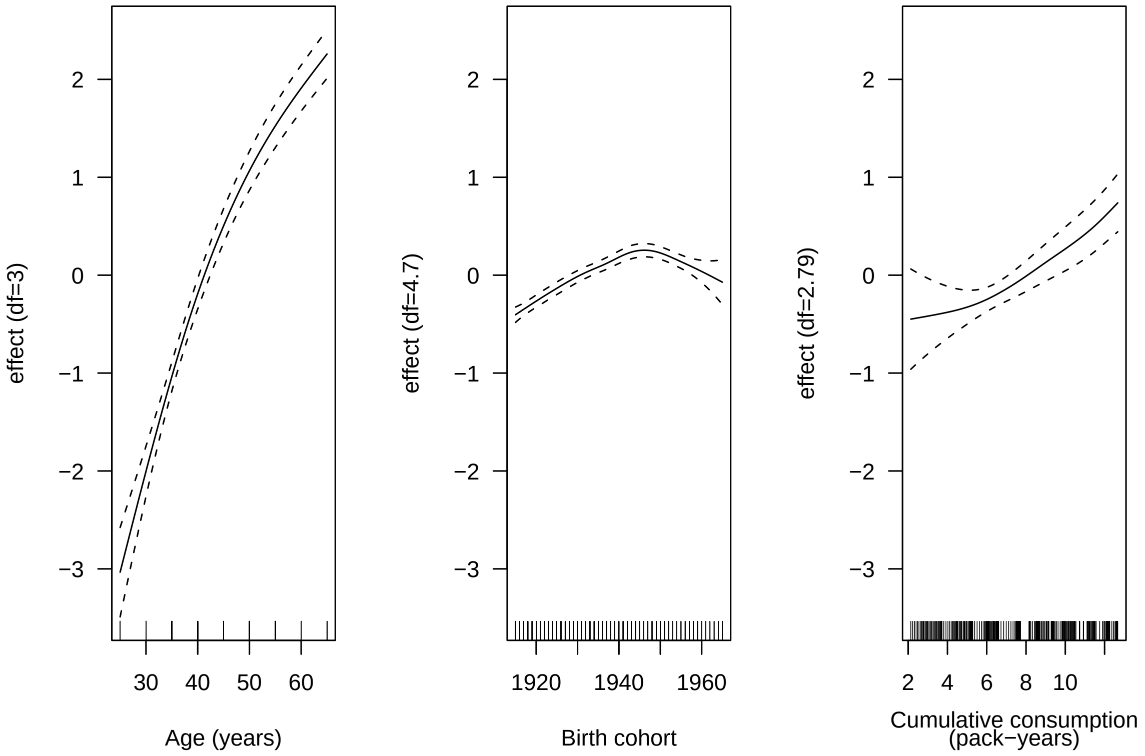


Figure 10.7: Smoothed covariate parameters for the GAM fit for Model A, females

"age + smoke" denotes the full model excluding any cohort terms. Any example of Model B can only be formally compared with other examples of Model B and the constant model.

For males, for both Model A and Model B, the models with the lowest Akaike's Information Criterion included a quadratic polynomial for cohort. From Table 10.4, smoking made a significant contribution to Model A. Moreover, for models with the same degrees of freedom, Model B performed well compared with models without smoking terms. The most parsimonious models based on parametric age were Model A with parametric age and Model B, and where both models included a quadratic polynomial for cohort. Model A provided a smaller Akaike's Information Criterion.

For females, the quadratic polynomial for cohort made a significant contribution to the model fitting (see Tables 10.3 and 10.4). In contrast to the males, Model A without cohort terms was not improved by the cumulative consumption parameter (for exclusion of the smoking term: $p = 0.27$ for non-parametric age, $p = 0.97$ for parametric age). The most parsimonious models based on parametric age were again Model A and Model B with quadratic polynomials for cohort. In contrast to males, Model B provided a smaller Akaike's Information Criterion, suggesting that Model

	Non-parametric age			Parametric age		
Model	df	Deviance	AIC	df	Deviance	AIC
Constant	330	131035.2	131037.2	330	131035.2	131037.2
Age	322	3948.5	3966.5	329	4595.5	4599.5
Age + cohort	321	2849.7	2869.7	328	3543.4	3549.4
Age + quadratic(cohort)	320	2573.6	2595.6	327	3117.3	3125.3
Model A						
Age + smoke	321	2346.7	2366.7	328	2610	2616
Age + cohort + smoke	320	2273.2	2295.2	327	2424.5	2432.5
Age + quadratic(cohort) + smoke	319	2238.0	2262.0	326	2379.4	2389.4
Model B						
Age + smoke	—	—	—	329	2637.0	2641.0
Age + cohort + smoke	—	—	—	328	2622.0	2628.0
Age + quadratic(cohort) + smoke	—	—	—	327	2440.2	2448.2

Table 10.2: Goodness of fit statistics for lung cancer models, males

B explained more variation for a given degrees of freedom.

	Non-parametric age			Parametric age		
Model	df	Deviance	AIC	df	Deviance	AIC
Constant	330	37018.8	37020.8	330	37018.8	37020.8
Age	322	2300.4	2318.4	329	2399.0	2403.0
Age + cohort	321	1995.8	2015.8	328	2082.6	2088.6
Age + quadratic(cohort)	320	1930.3	1952.3	327	1978.9	1986.9
Model A						
Age + smoke	321	2299.2	2319.2	328	2399.0	2405.0
Age + cohort + smoke	320	1960.7	1982.7	327	2041.3	2049.3
Age + quadratic(cohort) + smoke	319	1892.1	1916.1	326	1933.6	1943.6
Model B						
Age + smoke	—	—	—	329	2125.0	2129.0
Age + cohort + smoke	—	—	—	328	1967.1	1973.1
Age + quadratic(cohort) + smoke	—	—	—	327	1883.7	1891.7

Table 10.3: Goodness of fit statistics for lung cancer models, females

10.3.1 Rate equations

The fitted rate equations for Model A and Model B are shown for the models with quadratic polynomials for cohort.

Model A

The fitted equation for Model A for males was:

		Non-parametric age		Parametric age	
Model comparison	df	Males	Females	Males	Females
Model A					
Cohort | age	1	1098.8	304.6	1052.1	316.4
quadratic(cohort) | age	2	1374.9	370.1	1478.2	420.1
Smoke | age	1	1601.8	1.2	1985.5	0.0
Smoke | age + cohort	1	576.5	35.1	1118.9	41.3
Smoke | age + quadratic(cohort)	1	335.6	38.2	737.9	45.3
Cohort | age + smoke	1	73.5	338.5	185.5	357.7
quadratic(cohort) | age + smoke	2	108.7	407.1	230.6	465.4
Model B					
Cohort | age + smoke	1			15.0	157.9
quadratic(cohort) | age + smoke	2			196.8	241.3

Table 10.4: Change of deviance to assess contributions for smoking, period and cohort

$$\begin{aligned}\lambda &= \exp\big[-24.5 - 0.0003(\text{cohort} - 1950) - 0.00030(\text{cohort} - 1950)^2 + \\ &\qquad 0.0077\text{CumCons}\big]\,(\text{age} - 5)^{3.58} \qquad (10.10)\end{aligned}$$

Similarly, the fitted equation for females was:

$$\begin{aligned}\lambda &= \exp\big[-29.1 - 0.0049(\text{cohort} - 1950) - 0.00063(\text{cohort} - 1950)^2 + \\ &\qquad 0.0063 \cdot \text{CumCons}\big]\,(\text{age} - 5)^{5.01} \qquad (10.11)\end{aligned}$$

The fitted parameters are summarised in Table 10.5. The linear cohort term for males is not significant, while the similar term for females is small, suggesting that the quadratic polynomial for both sexes peaks close to the 1950 birth cohort. The other parameters are all highly significant.

For an indication of model stability, the correlation matrices for the fitted parameters are shown in Table 10.6. The estimates for the intercept and the age parameter are closely correlated, suggesting that the fitted parameters may not be stable. This may explain the large differences in point estimates between males and females.

Model B

For males, the fitted rate equation for Model B was:

	Males		Females	
Parameter	Estimate	(StdErr)	Estimate	(StdErr)
(Intercept)	-24.5	(0.3)	-29.1	(0.5)
cohort-1950	-0.00035	(0.00242)	-0.00488	(0.00241)
(cohort-1950)^2	-0.000297	(0.000045)	-0.000639	(0.000063)
CumCons	0.00773	(0.00029)	0.00627	(0.00094)
log(age-5)	3.58	(0.10)	5.01	(0.18)

Table 10.5: Estimates (and standard errors) for Model A with a quadratic polynomial for cohort, by sex

		(Intercept)	cohort-1950	(cohort-1950)^2	CumCons	log(age-5)
Males	(Intercept)	1.000	0.580	0.317	0.912	-0.994
	cohort-1950	0.581	1.000	0.901	0.742	-0.608
	(cohort-1950)^2	0.317	0.901	1.000	0.449	-0.331
	CumCons	0.912	0.742	0.449	1.000	-0.948
	log(age-5)	-0.994	-0.608	-0.331	-0.948	1.000
Females	(Intercept)	1.000	-0.045	-0.136	0.924	-0.996
	cohort-1950	-0.045	1.000	0.916	0.148	0.018
	(cohort-1950)^2	-0.136	0.916	1.000	0.010	0.121
	CumCons	0.924	0.148	0.010	1.000	-0.954
	log(age-5)	-0.996	0.018	0.121	-0.954	1.000

Table 10.6: Correlation matrices for the estimated parameters from Model A, by sex

$$\begin{aligned}\lambda &= 0.201 \times 10^{-12} \exp\Big[- 0.025(\text{cohort} - 1950) - 0.00058(\text{cohort} - 1950)^2 \Big] \\ &\times \Bigg\{ (\text{age} - 5)^{5.09} + \\ &\quad 10.99\, \pi_c\, \text{E}(\text{Dose}_c^*)\text{E}(\text{Dur}_c^*)^{5.01} \left[1 + 10.03 \frac{\text{var}(\text{Dur}_c^*)}{\text{E}(\text{Dur}_c^*)^2} \right] + \\ &\quad 10.99\, \pi_x\, \text{E}(\text{Dose}_x^*)\text{E}(\text{Dur}_x^*)^{5.01} \left[1 + 10.03 \frac{\text{var}(\text{Dur}_x^*)}{\text{E}(\text{Dur}_x^*)^2} \right] \Bigg\} \qquad (10.12)\end{aligned}$$

where Dose_c^* and Dur_c^* are the 5-year lagged estimates for dose and duration among current smokers, respectively, and Dose_x^* and Dur_x^* are the comparable estimates for former smokers. For females, the similar equation was:

$$\begin{aligned}
\lambda &= 0.981 \times 10^{-12} \exp\Big[-0.010(\text{cohort}-1950) - 0.00058(\text{cohort}-1950)^2 \Big] \\
&\quad \times \Bigg\{ (\text{age}-5)^{4.60} + \\
&\qquad 46.85\,\pi_c\,\text{E}(\text{Dose}^*_c)\text{E}(\text{Dur}^*_c)^{4.60}\left[1 + 8.27\frac{\text{var}(\text{Dur}^*_c)}{\text{E}(\text{Dur}^*_c)^2}\right] + \\
&\qquad 46.85\,\pi_x\,\text{E}(\text{Dose}^*_x)\text{E}(\text{Dur}^*_x)^{4.60}\left[1 + 8.27\frac{\text{var}(\text{Dur}^*_x)}{\text{E}(\text{Dur}^*_x)^2}\right] \Bigg\} \qquad (10.13)
\end{aligned}$$

Recall that ν_1 and ν_2 are the linear and quadratic terms for (cohort-1950), β is the log of the constant for current and former smokers and θ is the parameter for power for duration. The fitted parameters are shown in Table 10.7. All of the parameters are highly significant. The quadratic polynomial for cohort reaches a peak for male cohorts born close to 1930 and female cohorts born close to 1940.

	Males		Females	
Parameter	Estimate	(StdErr)	Estimate	(StdErr)
ν_1	-0.0252	(0.0018)	-0.0102	(0.0023)
ν_2	-0.000580	(0.000043)	-0.000580	(0.000063)
β	2.40	(0.09)	3.85	(0.19)
θ	5.01	(0.03)	4.60	(0.06)

Table 10.7: Estimates (and standard errors) for Model B with a quadratic polynomial for cohort, by sex

An investigation of the correlation matrices for the estimated parameters suggests that the fitted parameters between the constant term for smokers and the power term for duration are highly correlated. This may explain why the estimate for the constant term is higher and the power term is lower for females compared with the estimates for males.

10.3.2 Comparison between males and females

Modelling the rates with males and females together, the estimated rate ratio for females compared with males from Model B was 0.834 (95% confidence interval: 0.822, 0.850). Estimates of the rate ratios were relatively stable between models with and without cohort terms.

However the model that tested whether the constant term for smokers (β) and the power term for duration (θ) were different between males and females was a

Sex		$\hat{\nu}_1$	$\hat{\nu}_2$	$\hat{\beta}$	$\hat{\theta}$
Males	$\hat{\nu}_1$	1.000	0.955	0.012	0.192
	$\hat{\nu}_2$	0.955	1.000	0.059	0.111
	$\hat{\beta}$	0.011	0.059	1.000	-0.974
	$\hat{\theta}$	0.192	0.111	-0.974	1.000
Females	$\hat{\nu}_1$	1.000	0.939	-0.142	0.267
	$\hat{\nu}_2$	0.939	1.000	-0.097	0.190
	$\hat{\beta}$	-0.142	-0.097	1.000	-0.987
	$\hat{\theta}$	0.267	0.190	-0.987	1.000

Table 10.8: Correlation matrices for the estimated parameters from Model B, by sex

significantly better fit. The estimates suggested that the point estimates for females were higher for β and lower for θ. This suggests that either the biology is different between males and females or that the exposure data were different. One implication is that any fit using these data should be performed separately for males and females.

A model that included all three terms was not an appreciable improvement from the model with only the interaction terms for β and θ.

10.3.3 Projections

Projected smoking exposure

The projected prevalence of current smoking for those aged 35–69 years is shown in Figures 10.8 and 10.9. For both males and females, the middle projections would be expected to decline through the 1980s with a slower decline through the 1990s and then stability from around the year 2000. The validity of the middle projections, assuming constant transition rates from 1993, will be considered in the discussion. Relative to the middle projected estimates of prevalence, doubling and halving cessation would lead to appreciable changes in smoking prevalence. Moreover, zero uptake in smoking would be expected to cause a decline in prevalence for the age group 10–15 years after zero initiation began, after which prevalence would drop rapidly.

Projected lung cancer mortality

Middle projections The middle projections for lung cancer mortality were based on the middle smoking projections. For a comparison of the different models, see Figures 10.10 and 10.11. The confidence interval for Model A using a generalised

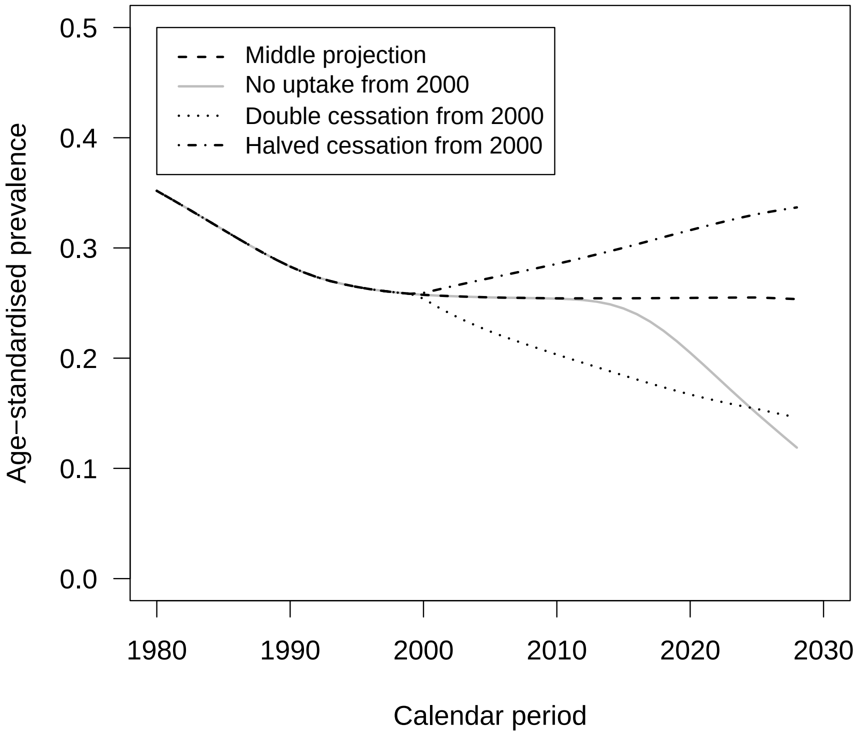


Figure 10.8: Projections for current smoking prevalence under different scenarios, Australian males aged 35-69 years, age-standardised to 1991 Australian population

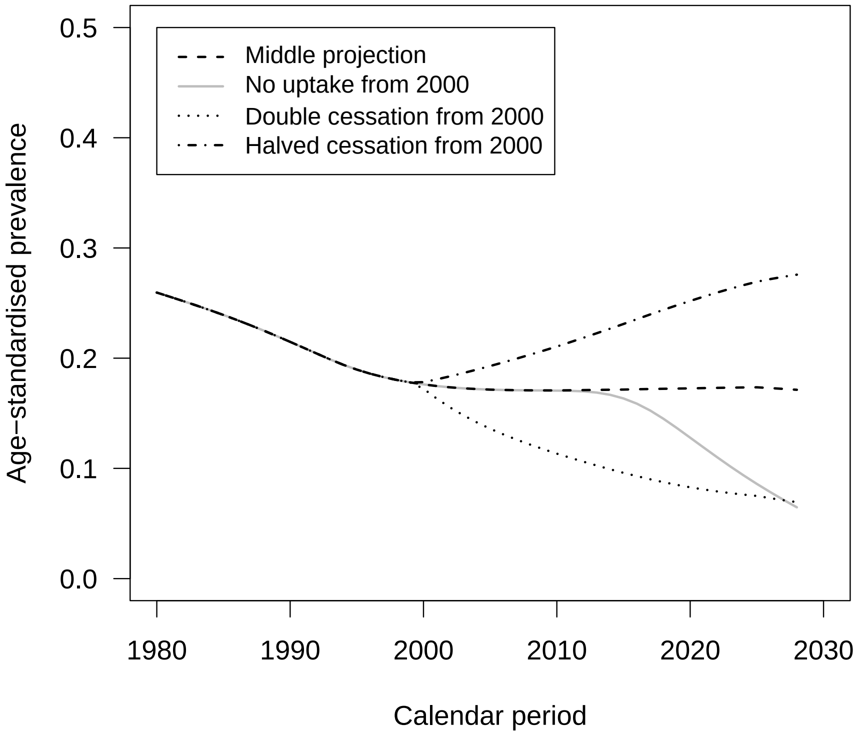


Figure 10.9: Projections for current smoking prevalence under different scenarios, Australian females aged 35-69 years, age-standardised to 1991 Australian population

linear model and Model B are shown. For males, the lung cancer rates for those aged 35-69 years would be expected to continue the rapid decline. The Model A projections based on generalised linear models were higher than those based on the generalised additive model. There was reasonable qualitative agreement between the generalised additive model and Model B. Moreover, the age-period model gave similar results to Model B.

For females, there was better qualitative agreement between all of the model fits (Figure 10.11). The general pattern was for a slow decline in lung cancer mortality rates for those aged 35–69 years. The confidence intervals for the generalised additive model included the means for the other three models. Model B would trace slightly lower than the other two estimates, although there was reasonable agreement with the age-period generalised additive model. Note that the scale for females is different from that for males.

Projections based on different smoking scenarios Projected lung cancer mortality rates were obtained from applying projected smoking exposure estimates to the fitted equations for Model B. For the given age group, zero uptake from the year 2000 would have negligible impact on lung cancer mortality to 2028, such that the projected mortality rates were almost indistinguishable from the middle projections (Figures 10.12 and 10.13) For males, the lung cancer rates would be expected to continue to decline under different smoking scenarios. The rapidity of such a decline would be sensitive to the cessation rate. Doubling or halving cessation would be expected to decrease or increase, respectively, the estimated middle projection at 2028 by approximately 30%.

The projected rapidity of decline of female lung cancer mortality rates for those aged 35–69 years would be sensitive to the different smoking scenarios (Figure 10.13). For a halving of the cessation rate relative to the middle projection from the year 2000, the lung cancer mortality rate would decline slowly. In contrast, a doubling of the cessation rate would give a considerably more rapid decline. Interestingly, the female lung cancer rates under the different scenarios were only slightly lower than the comparable rates for males. Moreover, the ratios of the lung cancer mortality rates at 2028 under the different smoking scenarios were expected to be roughly similar for males and females.

Estimates for the lung cancer mortality rates are shown in Table 10.9.

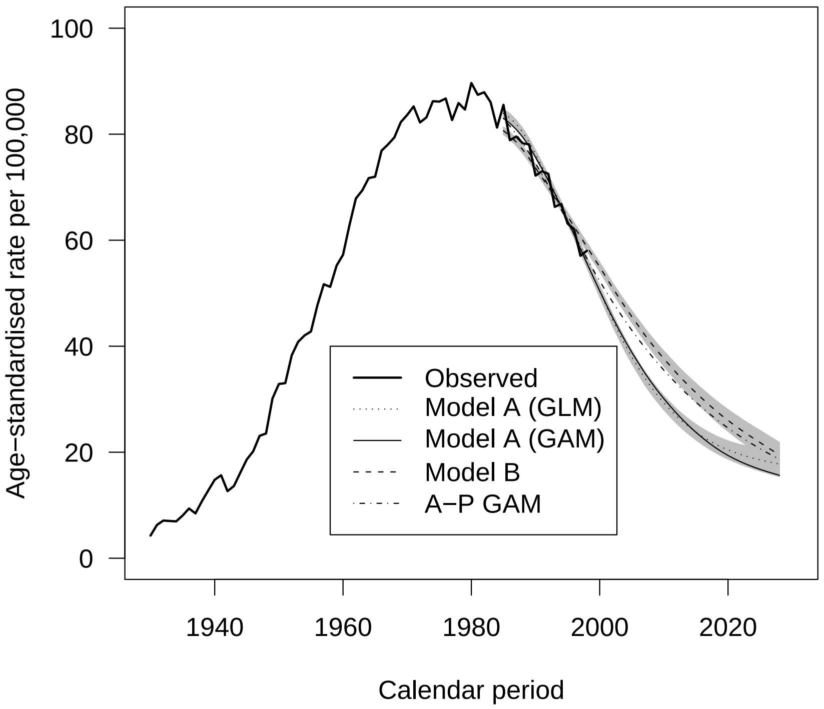


Figure 10.10: Middle projections for lung cancer mortality, Australian males aged 35-69 years, age-standardised to Segi's World population, for different models (grey shading = 95% confidence interval for mean for Model A (GLM) and Model B)

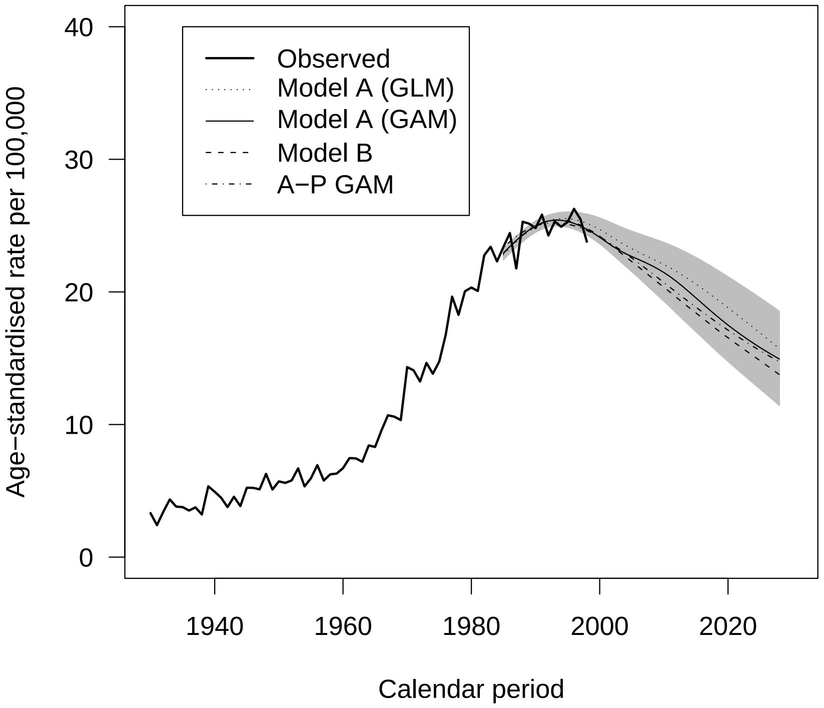


Figure 10.11: Middle projections for lung cancer mortality, Australian females aged 35-69 years, age-standardised to Segi's World population, for different models (grey shading = 95% confidence interval for mean for Model A (GLM) and Model B)

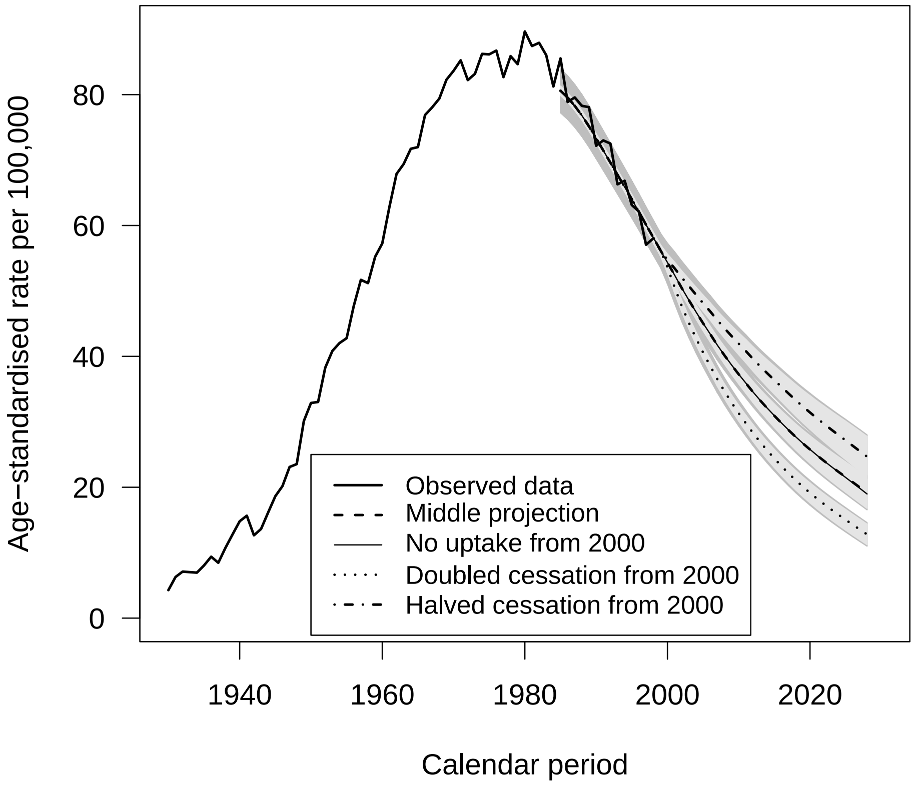


Figure 10.12: Projections for lung cancer mortality under different scenarios, Australian males aged 35-69 years, age-standardised to Segi's World population (darker grey shading = 95% prediction intervals for individuals; lighter grey shading = 95% confidence intervals for mean)

Table 10.9: Predicted age-specific lung cancer mortality rates per 100,000 per year, by sex, calendar period and by smoking scenarios, 2000–2024

			Age group (years)						
Sex	Scenario	Year	35-39	40-44	45-49	50-54	55-59	60-64	65-69
Males	Middle	2000-2004	1.3	4.0	9.4	19.2	34.3	55.5	84.5
		2005-2009	1.1	3.6	8.8	17.5	31.4	51.2	77.4
		2010-2014	1.0	3.1	8.1	16.3	28.5	46.3	70.7
		2015-2019	0.8	2.6	6.9	15.0	26.5	41.7	63.6
		2020-2024	0.6	2.1	5.8	12.8	24.3	38.7	57.2
	No uptake	2000-2004	1.3	4.0	9.4	19.2	34.3	55.5	84.5
		2005-2009	1.1	3.6	8.8	17.5	31.4	51.2	77.4
		2010-2014	1.0	3.1	8.1	16.3	28.5	46.4	70.7
		2015-2019	0.8	2.6	6.9	15.0	26.5	41.8	63.7
		2020-2024	0.4	1.9	5.7	12.8	24.3	38.8	57.3
	Cessation × 2	2000-2004	1.3	3.9	9.1	18.5	33.0	53.6	82.1
		2005-2009	1.1	3.4	8.1	15.8	28.3	46.3	70.9
		2010-2014	0.9	2.7	6.9	13.7	23.6	38.6	60.0

Table 10.9: (continued)

			Age group (years)						
Sex	Scenario	Year	35-39	40-44	45-49	50-54	55-59	60-64	65-69
		2015-2019	0.6	2.1	5.4	11.5	20.0	31.9	49.6
		2020-2024	0.5	1.5	4.1	8.9	16.7	26.9	41.2
	Cessation / 2	2000-2004	1.3	4.0	9.6	19.6	35.1	56.7	86.0
		2005-2009	1.2	3.8	9.3	18.8	33.9	55.0	82.5
		2010-2014	1.0	3.4	9.0	18.5	32.6	53.3	80.3
		2015-2019	0.9	3.0	8.0	17.9	32.3	51.4	78.0
		2020-2024	0.7	2.6	7.1	16.1	31.4	51.1	75.4
Females	Middle	2000-2004	1.3	4.0	9.4	19.2	34.3	55.5	84.5
		2005-2009	1.1	3.6	8.8	17.5	31.4	51.2	77.4
		2010-2014	1.0	3.1	8.1	16.3	28.5	46.3	70.7
		2015-2019	0.8	2.6	6.9	15.0	26.5	41.7	63.6
		2020-2024	0.6	2.1	5.8	12.8	24.3	38.7	57.2
	No uptake	2000-2004	1.3	4.0	9.4	19.2	34.3	55.5	84.5
		2005-2009	1.1	3.6	8.8	17.5	31.4	51.2	77.4
		2010-2014	1.0	3.1	8.1	16.3	28.5	46.4	70.7
		2015-2019	0.8	2.6	6.9	15.0	26.5	41.8	63.7
		2020-2024	0.4	1.9	5.7	12.8	24.3	38.8	57.3
	Cessation × 2	2000-2004	1.3	3.9	9.1	18.5	33.0	53.6	82.1
		2005-2009	1.1	3.4	8.1	15.8	28.3	46.3	70.9
		2010-2014	0.9	2.7	6.9	13.7	23.6	38.6	60.0
		2015-2019	0.6	2.1	5.4	11.5	20.0	31.9	49.6
		2020-2024	0.5	1.5	4.1	8.9	16.7	26.9	41.2
	Cessation / 2	2000-2004	1.3	4.0	9.6	19.6	35.1	56.7	86.0
		2005-2009	1.2	3.8	9.3	18.8	33.9	55.0	82.5
		2010-2014	1.0	3.4	9.0	18.5	32.6	53.3	80.3
		2015-2019	0.9	3.0	8.0	17.9	32.3	51.4	78.0
		2020-2024	0.7	2.6	7.1	16.1	31.4	51.1	75.4

10.4 Discussion

10.4.1 Summary

In summary, following a review of different lung cancer models, two sets of lung cancer regression models were fitted and used for projections. The first set of models included a simple measure of population smoking and was fitted with both generalised additive models and with generalised linear models. The second set of models included data on never smoker mortality rates, smoking prevalence, dose and dura-

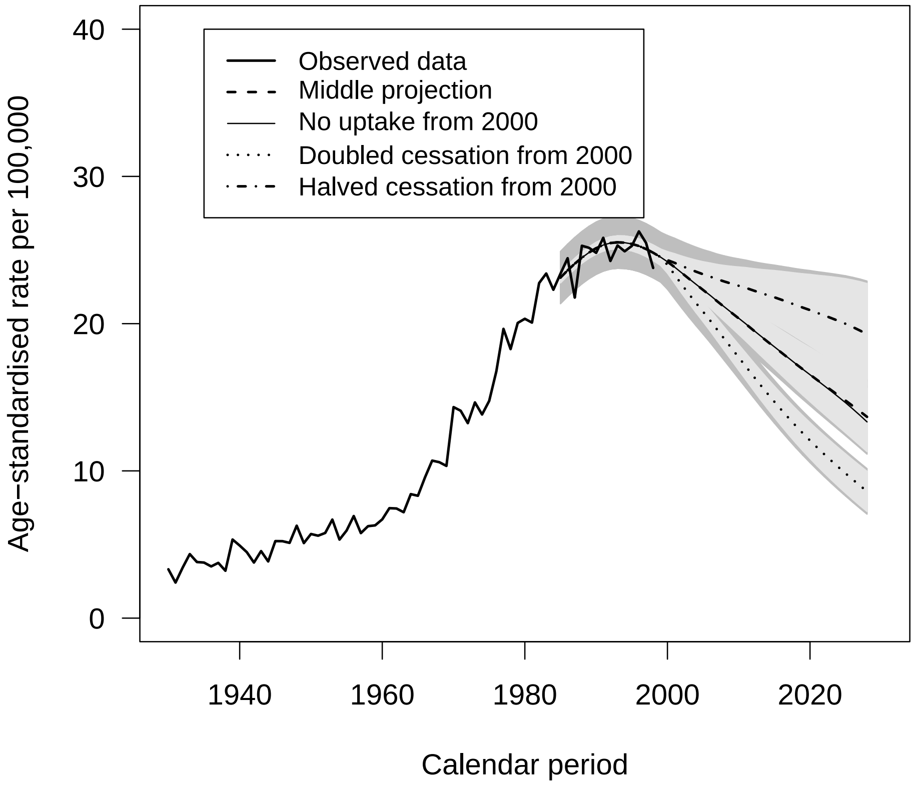


Figure 10.13: Projections for lung cancer mortality under different scenarios, Australian females aged 35-69 years, age-standardised to Segi's World population (darker grey shading = 95% prediction intervals for individuals; lighter grey shading = 95% confidence intervals for mean)

tion. The latter set of models was fitted using generalised non-linear models with variance estimated using the delta method and validated using the bootstrap.

Using a log link, Model A included a log term for (age-5), a quadratic polynomial for cohort and a linear term for cumulative consumption. Model B also included a quadratic polynomial for cohort.

Model A and Model B provided similar model performance. This has the advantage of allowing the choice of model to be determined by the available data, rather than being constrained by having one model being more valid than another.

Lung cancer projections for those aged 35–69 years from the different models were in reasonable agreement. For the middle projections of lung cancer rates, male rates were expected to continue to decline rapidly, while female rates were expected to decline more slowly. Lung cancer projections based on Model B were sensitive to changes in smoking cessation, while rate for the age group (35–69 years) were insensitive to changes in uptake out to 2028.

10.4.2 Limitations

There are a number of important limitations to these results. First, the validity of the regression modelling is limited by the validity of the smoking exposure data. This potential limitation has been the motivation for care in the analytical efforts of previous chapters. See Chapter 11 for further discussion.

Second, the lung cancer regression model may not be valid, where there is disagreement in the literature over the form of the lung cancer mortality rate for former smokers. It is also unclear whether the dose parameter should be linear or squared. Finally, a number of different models all described the data equally well (Thomas, 1988).

Third, as discussed in Section 10.1.3, the use of population-based data for modelling lung cancer rates from vitals has certain limitations. One of the strengths of using a multi-state smoking model is the availability of novel estimates, including measures of population heterogeneity. Exclusion of the variance terms for duration in Model B led to lung cancer mortality rate projections based on zero initiation being incorrectly higher than the middle projections.

Fourth, the variance estimates were crude because variance estimates for the covariates were not available. This is a major constraint of the current approach using multi-state models. Had the variance components been available, then the components could have been included in the delta-method (see Section 10.2.5). One consequence is that the confidence intervals for the means were conservatively narrow. One extension would be to provide confidence intervals for individual predictions, possibly using the bootstrap (Davison and Hinkley, 1997).

Fifth, projections suffer from the assumption that the future will be similar to the past (Hakulinen, 1996). Any major shift in behaviour would potentially invalidate any predictions.

The use of the generalised additive model with age and period does provide some inter-method validation that the middle projections may be valid.

A discussion of the limitations to the general approach will be given in Chapter 11.

10.4.3 Extensions

For fitting the lung cancer mortality rate regression models, the rates may be over-dispersed due to unexplained heterogeneity. The over-dispersion can be modelled using quasi-likelihood methods or using a negative binomial distribution (McCullagh and Nelder, 1989).

One potential extension would be to apply the smoking data to a two-mutation model for population-based lung cancer mortality. The two-mutation model has received a full theoretical development (Moolgavkar and Luebeck, 1990) and has been used to model population-based colon cancer (Moolgavkar and Luebeck, 1992).

Other extensions include: estimating the number of lung cancer deaths by applying the estimated rates to population projections; and modelling for a pulse intervention rather than a sustained change in cessation.

10.4.4 Interpretation

The exposure measures were carefully estimated and the different lung cancer regression models gave consistent results. However, given the limitations, care is recommended in the interpretation of the results.

All of the lung cancer regression models included terms for birth cohort. The terms represent variability that was not explained by either age or the given smoking parameters. The cohort terms can be explained either by aspects of tobacco exposure that were inaccurately measured, such as tar, or other sources of lung cancer risk, such as air pollution and industrial exposures.

The middle projections for current smoking prevalence were relatively stable from the late 1990s. This is at variance with information from the Commonwealth Department of Hea (1999) which suggested smoking prevalence declined appreciably through the late 1990s. It is *not* clear whether there has been a consistent time-series of smoking prevalence in Australia since 1995. The use of a middle projection of no change in prevalence would provide conservatively high estimates of exposure and of lung cancer mortality relative to prevalence that declines over time.

The public health message from the analysis is “Lung cancer mortality rates are expected to decline, however the speed of the decline is dependent upon the level of smoking cessation.”

Chapter 11

Conclusions

> The trouble with our times is that the future is not what it used to be.
> *Paul Valery* (1871–1945)

11.1 Introduction

A multi-state smoking model was developed and applied to observed smoking data from Australia and New Zealand. A variety of analytical approaches were used to derive estimates of smoking exposure adjusted for differential survival and changes in smoking behaviour. The estimates of smoking exposure were then used to model lung cancer mortality and to make projections.

It is proposed here that the use of a multi-state model with valid transition estimates provided a sound framework for modelling smoking as a risk factor for lung cancer mortality. If this proposal is well founded and the methods are made accessible for use in policy, then this research has the potential for making a substantial contribution to epidemiology.

The multi-state smoking models and its application to lung cancer mortality offered:

- A theoretical development of population-based smoking uptake and cessation for use by other researchers

- Estimators for smoking initiation and cessation based on current status and retrospective data

- A description of changing smoking dynamics

- A framework for Markov modelling of smoking behaviour, including a combination of current status and retrospective estimates
- Estimators of novel population-based smoking exposure parameters, including measures of population heterogeneity
- A framework for applying a multi-stage lung cancer model to estimate mortality rates given different smoking scenarios.

This final chapter discusses the limitations of the approach, outlines some possible extensions and discusses some implications.

11.2 Limitations

The use of population-based smoking exposure for modelling lung cancer mortality has inherent limitations (see Section 10.1.3). While individual-based studies are preferred on scientific grounds, they would in general be prohibitively expensive.

The following sections provide a general discussion of data limitations and estimation issues.

11.2.1 Data limitations

As described in Chapter 1, smoking exposure changed rapidly in Australia and New Zealand throughout the twentieth century. Aspects of particular relevance to the lung cancer modelling included: changing patterns of uptake and cessation; changes in the number of cigarettes smoked per smoker; the introduction of filters; changing tar yields from cigarettes; and other changes in cigarette composition and construction.

Limited data were available to describe the changes in exposure. Consistent current status prevalence data were only available from the 1970s for Australia and for New Zealand. Representative data on retrospective smoking behaviour were only available from the 1980s for Australia and from the 1990s for New Zealand for cohorts born from 1910. Data for total population tobacco consumption, population estimates and mortality were available for most of the twentieth century.

Given these data restrictions, the approach taken was to restrict analysis to a limited section of the Lexis diagram (see Figure 10.3 on page 224). This has the advantage of making few assumptions about the nature of smoking for imprecisely observed or unobserved cohorts. Moreover, the later birth cohorts were more homogeneous with respect to the use of filters. There was also no need to adjust for possible under-diagnosis of lung cancer as a cause of deaths in earlier cohorts.

The main disadvantage of performing an analysis over the limited section of the Lexis diagram was that a small number of deaths were observed at older ages. The analysis included few deaths for those aged 65–69 years. Ideally older ages would have been included, however the aetiology of smoking cessation at older ages is even less understood than cessation at younger ages (Enstrom, 1999). As a consequence, projections were restricted to premature mortality, while projections have not been provided for lung cancer mortality across all ages.

Even given the restriction on the Lexis diagram, some assumptions about unobserved cohorts were unavoidable. Most notably, prevalence of current smoking was required back to 1950 for calculation of dose for earlier birth cohorts at younger ages. There was some disagreement between estimated and observed smoking prevalence, such that the estimated dose for those cohorts may be biased (see Chapter 9).

11.2.2 Estimation issues

One implication of using a narrow section of the Lexis diagram was that parameter estimates were based on smaller numbers. This imprecision was most noticeable for the non-parametric estimates of the cessation rates at the boundaries of the Lexis diagram. The methods used were validated against a simulation data set, however small variations at the boundaries led to unstable cessation estimates. The proportionality of the hazards was questionable for the older cohorts, however a parametric or semi-parametric approach may provide more reliable estimates for younger cohorts.

With increasing model complexity, variance estimation became more difficult. Variance estimates were available for the retrospective transition rates. Variance estimates for the current status and census-based transition rate estimates could also be estimated using bootstrapping, Monte Carlo simulation or Bayesian methods (Ogata et al., 2000), but these methods have not been implemented here. Matrix methods are also available for variance estimation of the transition probabilities but these methods have not been implemented here (Andersen et al., 1993). An alternative approach would be to implement a full Bayesian analysis, such as has been used for HIV/AIDS modelling (De Angelis et al., 1998).

11.3 Research extensions

The variety of estimates from the multi-state model provide opportunities for applications to a range of public health questions, including:

1. Historical cross-sectional survey data
2. Assessment of modified attributable risk estimates
3. Lung cancer incidence and survival
4. Application of the methods to other conditions.

11.3.1 Historical cross-sectional survey data

Roy Morgan Research, whose surveys are used by the Anti-Cancer Council of Victoria to report national smoking prevalence, has indicated that it has archival material on current smoking prevalence. They have expressed the possibility that the archives could be investigated.

11.3.2 Assessment of modified attributable risk estimates

From a methodological perspective, the projections can be taken as a standard for use in evaluating the validity of projected changes in smoking attributable mortality based on modified attributable risk estimates or *generalised impact fractions* or *potential impact fractions* (Morgenstern and Bursic, 1982; Gunning-Schepers and Barendregt, 1992). The proposed analysis would be similar in intent to the evaluation of the `Prevent` model using micro-simulation (Brønnum-Hansen, 1999), with a greater emphasis on valid modelling of the lung cancer aetiology.

11.3.3 Lung cancer incidence and survival

It could be possible to extend the multi-state model to include intermediate health states and transitions, such as incident lung cancer cases and relative survival. This would be dependent upon developing appropriate risk functions for survival, using national survival data currently being analysed at the Australian Institute of Health and Welfare. Given state and international variation in lung cancer survival, this may provide a mechanism to assess the potential effect of improved care for lung cancer patients.

11.3.4 Application of the methods to other conditions

Development of risk functions for other sources of tobacco attributable burden of disease, such as chronic obstructive pulmonary disease and upper aero-digestive cancers, could permit similar predictive modelling. For any of these, however, the data that are needed would probably be as or more limited than data available for modelling lung cancer.

11.4 Implications

Smoking continues to have an enormous impact on the health of Australians (Mathers et al., 1999). Given the costs to the community of tobacco-related disability and costs of tobacco control programs to tax payer, there is a strong need to quantify the impact of tobacco interventions on the burden of disease.

11.4.1 Modelling as multi-disciplinary research

To assess how these methods may be used in the future, it may be useful to consider how similar methods have been received in the past. Tolley et al. (1991) used a closely related model, however their results and methods have received little attention.

One possible explanation for the limited attention may be the level of methodological sophistication of multi-state models. Although multi-state models are a direct extension of familiar epidemiological models, there remains a technical hurdle for the use of such models. Teaching of such methods may encourage their broader use. Alternatively, multi-state models can be used to validate simpler models, such as proposed in Section 11.3.1.

An alternative explanation is that Tolley et al. (1991) paid limited attention to parameter estimation. Given the complexity of the multi-state models, it is questionable to not use comparable effort on parameter estimation.

The multi-state models provide powerful tools, however there are greater demands on those developing public health models to ensure that model inputs are valid (Gunning-Schepers, 1999). The given research question required knowledge of:

1. Epidemiology, for a valid representation of the aetiology
2. Mathematical models, for realistic and useful models
3. Statistics, for valid estimation of parameters and their precision.

This emphasises the need for a multi-disciplinary approach, where valid modelling cannot proceed with expert input from different disciplines. This relates to the recent discussion on re-defining *modern epidemiology*, where commentators suggested that epidemiology cannot address many important public health questions without a multi-disciplinary approach (Wall, 1995; Pearce, 1996).

In reality, it is often easier for investigators to remain isolated within their discipline. Further encouragement may be required to bring multidisciplinary teams together.

11.4.2 Practical use of the methods

For these methods to be used more widely, some refinement of the methods would be required (see Section 11.3). Simplifications of the methods may also be warranted, such as hazard estimation being based on only retrospective data using a validated estimation method. The incorporation of the methods into other population health models would help to improve the validity of such models.

The data used for analysis could be improved further. Existing data sources may be under-utilised, such as suggested in Section 11.3.1. For future data collection, the incorporation of standardised questions on smoking uptake and cessation will provide important exposure data for future generations.

Periodic repetition of the projections will help to assess the validity of the rate equations and to estimate lung cancer mortality under changing exposure distributions. Such efforts would need to be performed only once every 5–10 years.

The methods can be used to help improve tobacco control planning. One such application would be to estimate changes in cessation rates over time, which could be used to assess whether tobacco control activities have helped to sustain or improve quitting. The shift from cross-sectional measures of smoking to dynamic measures will improve the representation of changing behaviour. The lung cancer incidence rate projections could be used to improve health care planning by applying predicted rates to population projections to estimate the number of new cases.

11.4.3 Lung cancer mortality projections

Qualitatively, there was good evidence that lung cancer mortality rates for those aged 35–69 years would decline among males and females over 2000–2028. The direction of change was largely independent of any predictable future smoking behaviour. However the rate of decline of female lung cancer mortality rates and the absolute level of lung cancer mortality rates at 2028 for both sexes were very sensitive to smoking cessation rates.

These results are in qualitative agreement with projections based on cumulative consumption reported to the Australian Commonwealth Government (Appendix A). The Commonwealth report presented annual numbers of lung cancer deaths aged 30–74 years, suggesting that the number of female lung cancer deaths would increase during 2000–2010 due to increasing population size.

As a general premise, projections should be undertaken with care. The ultimate judgement of any projection is the realisation of the future, hence the validity of the projections presented here can be assessed in the next 10–20 years.

11.4.4 Public health

The good news is that tobacco control efforts in the 1960s through to the present have prevented many premature lung cancer deaths. However the number of lung cancer deaths in the future is dependent upon the smoking behaviour of the population today and tomorrow. The only certain way to avoid lung cancer deaths over the next 30 years is through increased smoking cessation.

Appendix A

Report to the Commonwealth: Contribution of decline in tobacco smoking to the decline in lung cancer mortality in Australia

Mark Clements
Richard Taylor
Department of Public Health and Community Medicine
Faculty of Medicine, University of Sydney [1] [2]

A.1 Introduction

Mortality from lung cancer rose in many countries from the first half of the twentieth century [Stanley et al. 1988; Vakil 1988; Thom, Epstein 1994], including in Australia [Rohan, Christie 1980]. Whereas lung cancer mortality has stabilised or started to decline in males by the late twentieth century, in many countries, it continues to rise in women as a manifestation of the later commencement of the epidemic [Vutuc, Gredler 1986; Boyle, Roberstson 1987; La Vecchia et al. 1988; Lee et al. 1990; Zheng et al. 1994; Janssen-Heijnen et al. 1995; Kubik, Plesko 1998; Nordlund 1998].

This pattern has also been observed in Australia [Giles et al. 1991]. In the 25–44 year age range Tasmanian women have double the lung cancer incidence rate than men and smoking is higher in young women than young men [Dwyer et al. 1994].

[1] Acknowledgements: Robert Gibberd and Eileen Doyle are thanked for their assistance.

[2] This report has been reformatted for inclusion in the thesis. References are cited using square brackets [] and provided on pages 273– 278.

For 1992, English and colleagues estimated that a high proportion of lung cancer mortality could also be attributed to smoking, with estimates of 84% for males and 77% for females [English et al, 1995]

Lung cancer is a highly fatal condition and thus mortality reflects incidence. Thus Australian lung cancer mortality data which are available from the first half of the twentieth century (from death registration) can be used to assess trends over a long period; lung cancer incidence is only available from NSW from 1972, and Australia-wide from 1980 (from cancer registries). Tobacco consumption data for Australia exist in various forms back to the early years of the 20th century, although estimates of age and sex specific consumption in early periods require a combination of data and assumptions.

The method adopted here to estimate the contribution in the decline in tobacco smoking to the decline in lung cancer mortality in Australia involves the development of a statistical model relating data on lung cancer mortality and tobacco smoking during the 20th century. Comparisons are then made between actual lung cancer mortality and that predicted if tobacco consumption did not decline but remained at its Australian zenith, remains at current consumption, and declines at various rates.

Other approaches to the attribution of declines in lung cancer mortality to declines in tobacco consumption are also discussed.

Exposure estimates were based around methods developed by Todd [1978] and applied to Australia by Doyle [1985]. Mortality and population data were taken from Australian Bureau of Statistics (ABS) statistics, and cumulative exposure estimates were derived from tobacco consumption data and tobacco smoking prevalence estimates. Risk models were fitted for lung cancer mortality rates against explanatory variables including age, period and cumulative exposure to cigarette smoking [Stevens, Moolgavkar 1984]. Interventions were modelled by possible changes to cigarette consumption.

A.2 Methods

A.2.1 Lung cancer mortality

Lung cancer mortality and population data were obtained from the Australian Bureau of Statistics for 1931–1995. Data were available by five year age groups and by sex for ages 30–74 years, which is the truncated age group used in an analysis of cancer trends by the International Agency for Research on Cancer (IARC). The numbers of lung cancer deaths under age 30 years were small; approximately 0.1% of all lung cancer deaths.

The coding of lung cancer mortality in Australia has changed over time. The definition for lung cancer used was "malignant neoplasm of the lung, bronchus and trachea" from 1968 [Coleman et al, 1993]. The specific codes for the different revisions of the International List of Causes of Death (1931–1949) and the International Classification of Disease (1950–1998) are given in Table A.1. For the period 1931–1949, the inclusion of malignant neoplasms of the pleura and mediastinum and the exclusion of trachea involve very small numbers. For the period 1950–1967, the inclusion of malignant neoplasm of the pleura also involves very small numbers. With exception of a rise in the incidence of cancer of the pleura for the 1980s and 1990s, following an increase in mesothelioma, lung and bronchus have accounted for virtually all of malignant neoplasms of the trachea, bronchus, lung, pleura and mediastinum.

Revision	Period	Code	Classification terms
Fourth	1931–1939	47b	Lung and pleura, which includes bronchus and mediastinum
Fifth	1940–1949	47b	Lung and pleura, which includes bronchus and mediastinum
Sixth	1950–1957	162–163	Trachea, bronchus and lung, which includes pleura
Seventh	1958–1967	162–163	Trachea, bronchus and lung, which includes pleura
Eighth	1968–1978	162	Trachea, bronchus and lung
Ninth	1979–1998	162	Trachea, bronchus and lung

Table A.1: Coding of malignant neoplasm of the lung, bronchus and trachea

Adjustment was made for under-diagnosis of lung cancer mortality during early years using estimates used by Mantel et al. [1986] (see Table A.2).

Quinquennia	Diagnostic accuracy
1931–1935	40
1936–1940	45
1941–1945	55
1946–1950	69
1951–1955	75
1956–1960	80
1961–1965	90
1966–1995	100

Table A.2: Indices of diagnostic accuracy of recorded lung cancer deaths in England and Wales (values reproduced from Table 3.2 of Doyle [1985] with the permission of E. J. Doyle)

For longer term patterns of lung cancer, estimates were available for cancer of the respiratory system and intra-thoracic organs (ICD-9 codes 162–165). Data for

1910–1984 were extracted from Holman et al (1982, 1987), and data for 1985–1994 were obtained from the ABS.

A.2.2 Estimation of historical cumulative consumption of cigarettes by birth cohorts

Doyle [1985] provides estimates of consumption of cigarettes per day *per person* by sex and age for the quinquennia 1921–1925 through to 1971–1975. Updated consumption data were derived for the quinquennia 1976–1980 through 1991–1995. Results were available by quinary quinquennia (5 year age group by 5 year period) through to ages 70–74 years.

Estimates were based on total cigarette consumption derived from tobacco consumption data, which was then apportioned for each period based on the estimated proportion of consumption by sex and by age group. The estimated tobacco consumption was then expressed as cigarettes per day *per person.*

Total cigarette consumption was based on excise and customs duty data for manufactured cigarettes and loose tobacco as provided by the Australian Bureau of Statistics. Following Doyle [1985], it was assumed that 76% of loose tobacco will be used for hand-rolled cigarettes, and that an average 1226 cigarettes would have been made from a kilogram of loose tobacco.

For the period 1921–1975, the estimated proportion of average consumption per day *per person* by sex and age was based on average consumption per day *per person* data for 1972–1979, and assumptions about the historical smoking by females between 1921–1970. Female smoking was assumed to follow an “S” curve from a low level in the 1930s to the levels observed by recent surveys. The relative consumption per day *per person* for each age group by sex was assumed constant for the period 1921–1970 at the level for 1972/73.

For the period 1976–1995, the relative consumption per day *per smoker* for each age group by sex was assumed constant at the level for 1978/79. This assumption was supported by the observation that there was no trend in the age- and sex-specific average consumption per day *per smoker* for 1972–1978. Moreover, estimates for the average consumption rate for 1980 were not markedly different (Hill and Gray, 1982). Prevalence estimates for 1976–1995 (Gray and Hill, 1975; Gray and Hill, 1977; Hill and Gray, 1982; Hill and Gray, 1984; Hill et al, 1988; Hill et al, 1991; Hill and White, 1995; Hill et al, 1998) were multiplied by the average consumption per day *per smoker* to estimate the average consumption per day *per person.*

The effect due to differential survival from smoking was corrected by using estimates from Australian life tables and assuming an all-cause mortality rate ratio of 1.50 for smokers versus non-smokers.

Cumulative exposure was estimated by taking the sum of the tobacco consumption estimates across age groups within a birth cohort up to the end of each sequential age bracket.

A.2.3 Descriptive estimates and risk estimation models

To describe the period changes in cigarette consumption and lung cancer mortality, rates for those aged 30–74 years were age-standardised to the 1991 Australian population.

For risk estimation, Poisson models were used that included cumulative consumption. The age-period-consumption model for of the lung cancer mortality rate λ_{ijk} for age i, period j and cohort k was

$$\log(\lambda_{ijk}) = \mu + \alpha_i + \beta X_{ijk} + \gamma_j,$$

where μ is an intercept term, α_i and γ_j are the corresponding effects due to age and period, and β is the estimate for the effect due to cumulative consumption X_{ijk}.

Other models were investigated but proved less satisfactory, including the probit model of Brown and Forbes [1974]. The models were fitted using generalised linear models using the `genmod` procedure in `SAS` [SAS Institute Inc, 1996]. The Poisson model used a log link and Poisson error distribution. Residual deviance was used as a measure for the goodness of fit. To investigate the role of the different variables, change in deviance statistics are reported for comparisons of the model with the specific variable excluded with the full model.

A.2.4 Tobacco consumption scenarios

Different scenarios were considered for cigarette consumption rates. The scenarios are described in Table A.3. The rate of decline in cigarette consumption rate experienced from 1971–1995 was estimated using an age-period model of the cigarette consumption rate c_{ijk} for age i, period j and cohort k:

$$\log(c_{ijk}) = \mu + \alpha_i + \gamma\, k,$$

where γ is the average period rate of decline across all ages. The hypothesised cigarette consumption rates were applied to cohorts from 1971–1975 for the worst

Scenario	Assumed age-sex specific cigarette consumption rates
Worst	Remains at 1971–1975 levels
Current	Remains at 1991–1995 levels
Slow decline	One percent rate of decline from 1991–1995 level
Fast decline	Rate of decline experienced during 1971–1995 from 1991–1995 level

Table A.3: Different scenarios for cigarette consumption rates

scenario and to cohorts from 1991–1995 for the other scenarios out to 2010 to estimate cumulative consumption.

The risk functions were then applied to the different cumulative cigarette consumption distributions to estimate lung cancer mortality rates and numbers. Age and period effects from the risk model were assumed to be at the levels estimated for periods before 1991–1995, and then stable at the 1991–1995 level after the 1991–1995 quinquennium. Given that the age, period and cumulative consumption were fully specified, the lung cancer mortality rate could be predicted. The rates were applied to Series II population projections from the ABS to predict the number of lung cancer deaths.

A.3 Results

A.3.1 Period changes in tobacco consumption and lung cancer mortality

The changes in consumption of tobacco products over the past century are shown in Figure A.1. The pattern of the rise and then fall of consumption is well recognised. These data were used as one of the primary inputs to the estimation of cumulative cigarette consumption.

A period analysis of cigarette consumption and lung cancer mortality shows that lung cancer mortality followed the rise in cigarette consumption by 10–30 years (Figure A.2). Female lung cancer mortality rates began to rise approximately 30 years after consumption began to rise. Male lung cancer rates began to fall about 10 years after consumption began to fall, however, female consumption has been falling for 15 years without a commensurate decline in lung cancer rates.

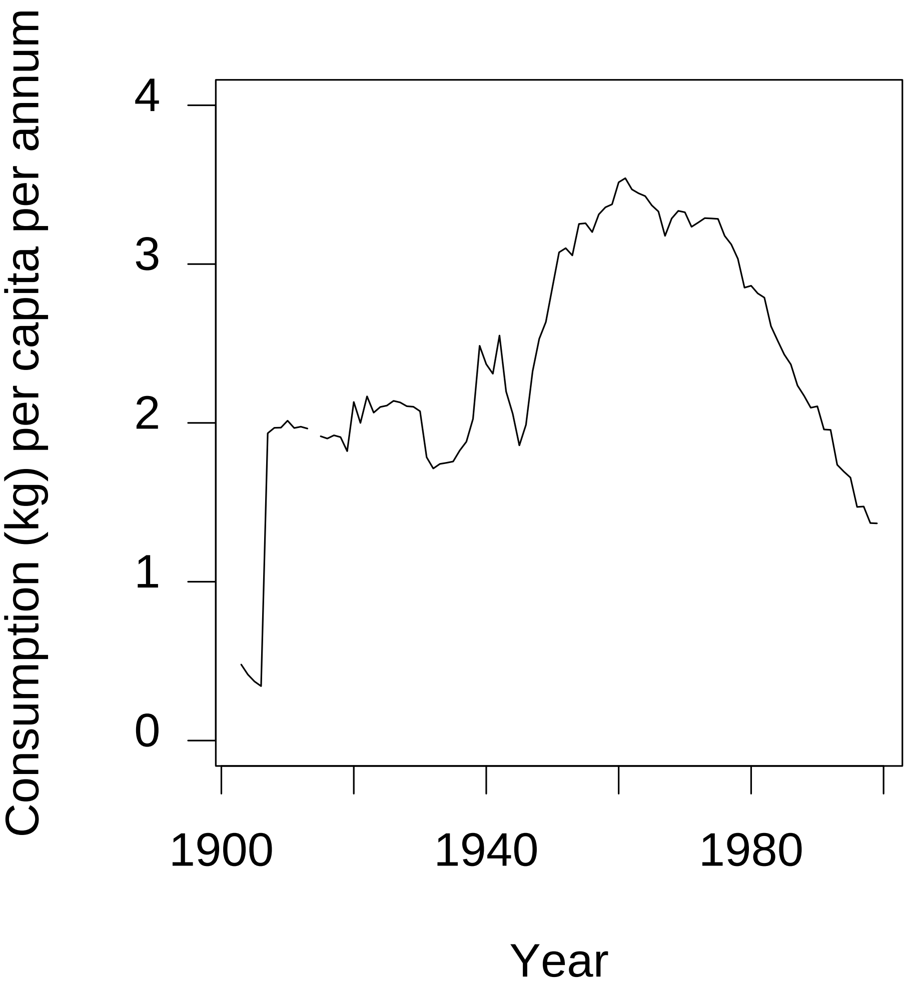


Figure A.1: Consumption of tobacco products in Australia, 1903–1999

A.3.2 Period and birth cohort trends in lung cancer mortality

To consider cohort and period effects jointly, mortality rates for cancer of the respiratory system and intra-thoracic organs are shown for males and females in Figures A.3 and A.4, respectively. The general pattern is for a rise with age, however there are changes across time, whether measured by cohort (on the diagonal) or period (on the horizontal). One summary interpretation of the figures is that the period and cohort effects can not easily be separated.

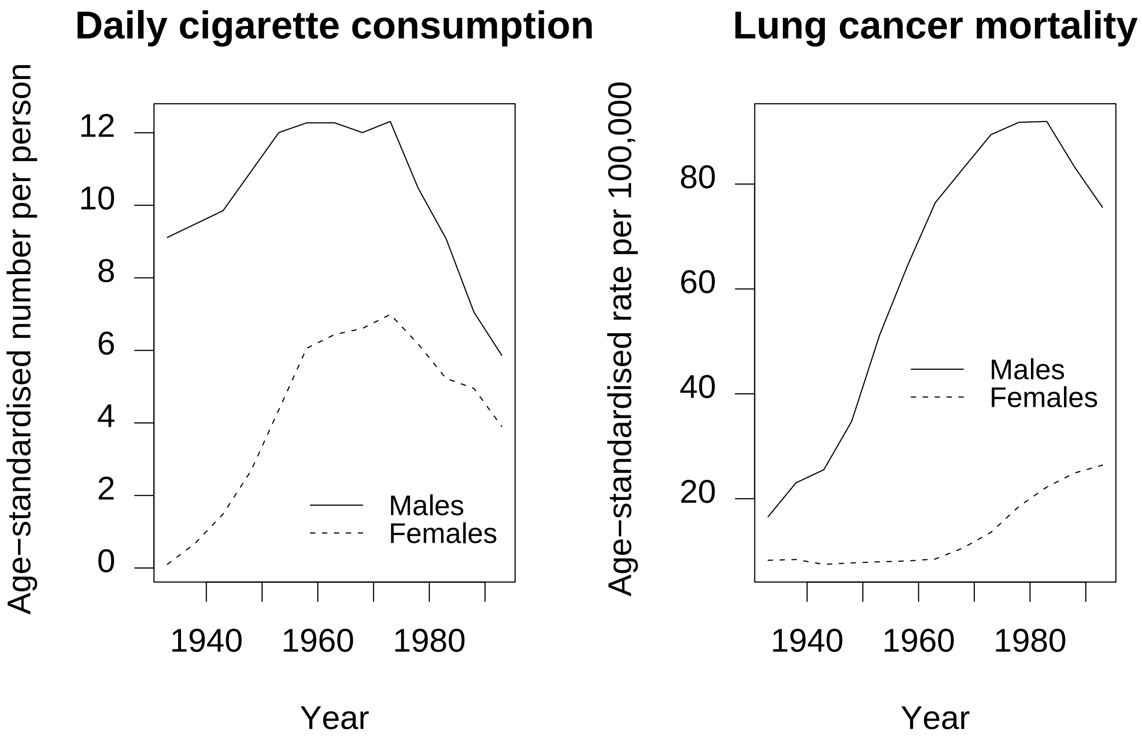


Figure A.2: Period effects of daily cigarette consumption and lung cancer mortality rates, ages 30–74 years, by sex and quinquennium, 1933–1993

A.3.3 Model of tobacco consumption and lung cancer mortality

The model fits for the age-period-consumption models are shown in Table A.4. The χ^2 for exclusion of age suggested that age is an important explanatory variable for lung cancer, consistent with the literature. Moreover, the consumption effect parameter consistently improved the models.

			Change in deviance for exclusion of variable		
Model	Sex	Deviance	Age	Period	Consumption
Poisson	Male	114.4	2231.0	536.9	263.9
	Female	131.6	17942.1	146.2	263.3
	(df)	(59)	(8)	(12)	(1)

Table A.4: Age-period-consumption model fits for lung cancer mortality, ages 30–74 years

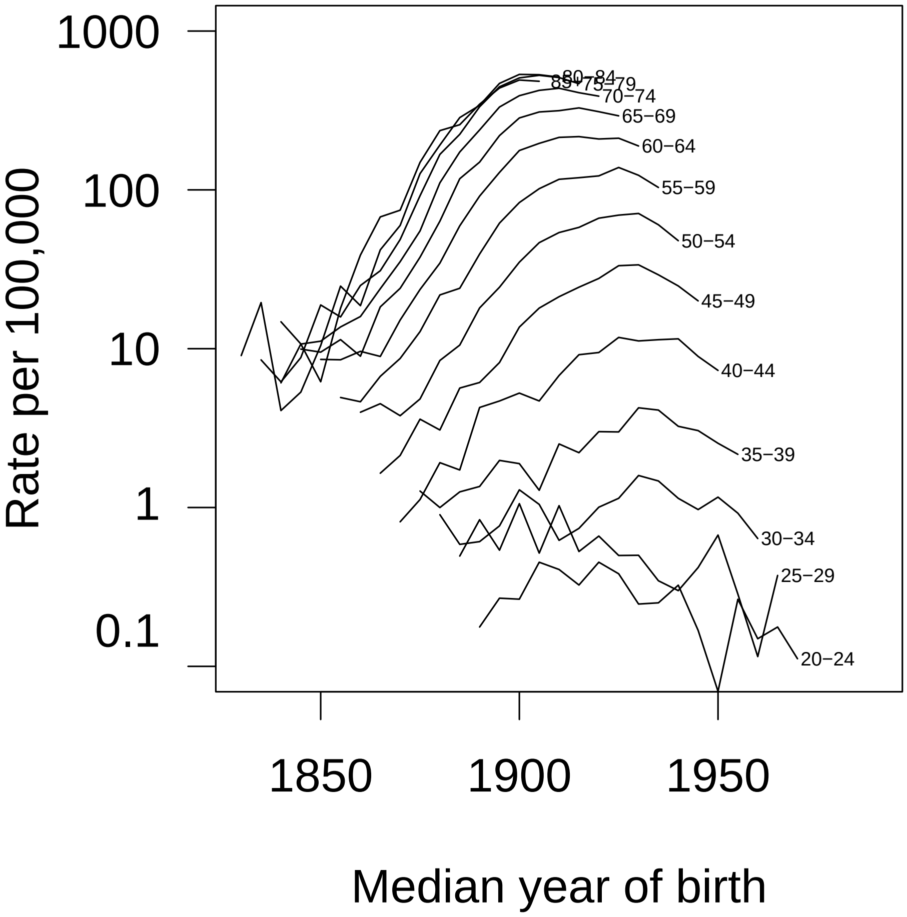


Figure A.3: Respiratory and intra-thoracic cancer mortality rates per 100,000, males

A.3.4 Alternative tobacco consumption scenarios

The alternative tobacco consumption scenarios for males are shown in Figure A.5. Scenarios for females are similar. The results for the lung cancer mortality rate estimations projections are shown for males and females in Figure A.6 and A.7, respectively. By applying the rate projections to population projections, estimates of the number of projected lung cancer deaths were obtained (see Figure A.8 and A.9). The area between the rate under the worst scenario and the other scenarios represents the avoided mortality, suggesting that large numbers of deaths have been avoided by smoking policies. The numbers of deaths are tabulated by the different scenarios in Table A.5, and the deaths averted in Table A.6.

Over 1971–1995 the cumulated number of deaths averted from lung cancer from the decline in tobacco smoking was 19,700 for men and 3,600 for women.

	Males					Females				
Quin-				Slow	Fast				Slow	Fast
quennium	Actual	Worst	Current	decline	decline	Actual	Worst	Current	decline	decline
1931–1935	94	-	-	-	-	45	-	-	-	-
1936–1940	160	-	-	-	-	58	-	-	-	-
1941–1945	235	-	-	-	-	69	-	-	-	-
1946–1950	438	-	-	-	-	101	-	-	-	-
1951–1955	762	-	-	-	-	126	-	-	-	-
1956–1960	1113	-	-	-	-	152	-	-	-	-
1961–1965	1591	-	-	-	-	199	-	-	-	-
1966–1970	2067	-	-	-	-	301	-	-	-	-
1971–1975	2543	2543	-	-	-	434	434	-	-	-
1976–1980	2900	3070	-	-	-	649	642	-	-	-
1981–1985	3232	3702	-	-	-	862	924	-	-	-
1986–1990	3236	4458	-	-	-	1051	1267	-	-	-
1991–1995	3235	5304	3235	3235	3235	1214	1668	1214	1214	1214
1996–2000	-	6131	2901	2901	2901	-	2093	1325	1325	1325
2001–2005	-	6971	2523	2502	2474	-	2557	1397	1389	1368
2006–2010	-	8097	2271	2214	2142	-	3141	1493	1465	1402

Table A.5: Observed and projected annual number of lung cancer deaths, ages 30–74 years

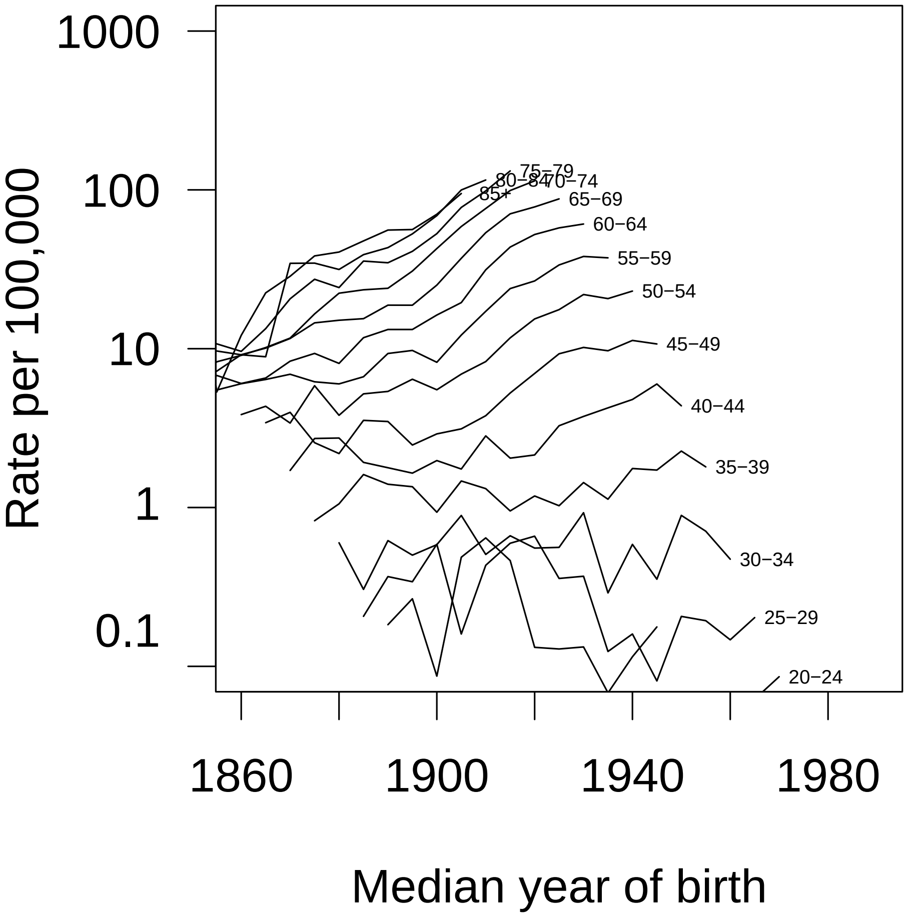


Figure A.4: Respiratory and intra-thoracic cancer mortality rates per 100,000, females

A.4 Discussion

A.4.1 Approaches to estimating the effect of reduction in tobacco smoking on decline in lung cancer mortality

Approaches which have been employed to estimate the contribution of reduction in tobacco smoking to the decline in lung cancer mortality in populations can be divided into those which use data on tobacco consumption and those which do not. Methods which do not use tobacco consumption basically assume that declines in lung cancer mortality are a consequence predominately of reduction in tobacco smoking, and that contributions from occupational and other environmental exposures are minimal.

Approaches which do not require data on tobacco consumption

(a) Projection of period and birth cohort rates

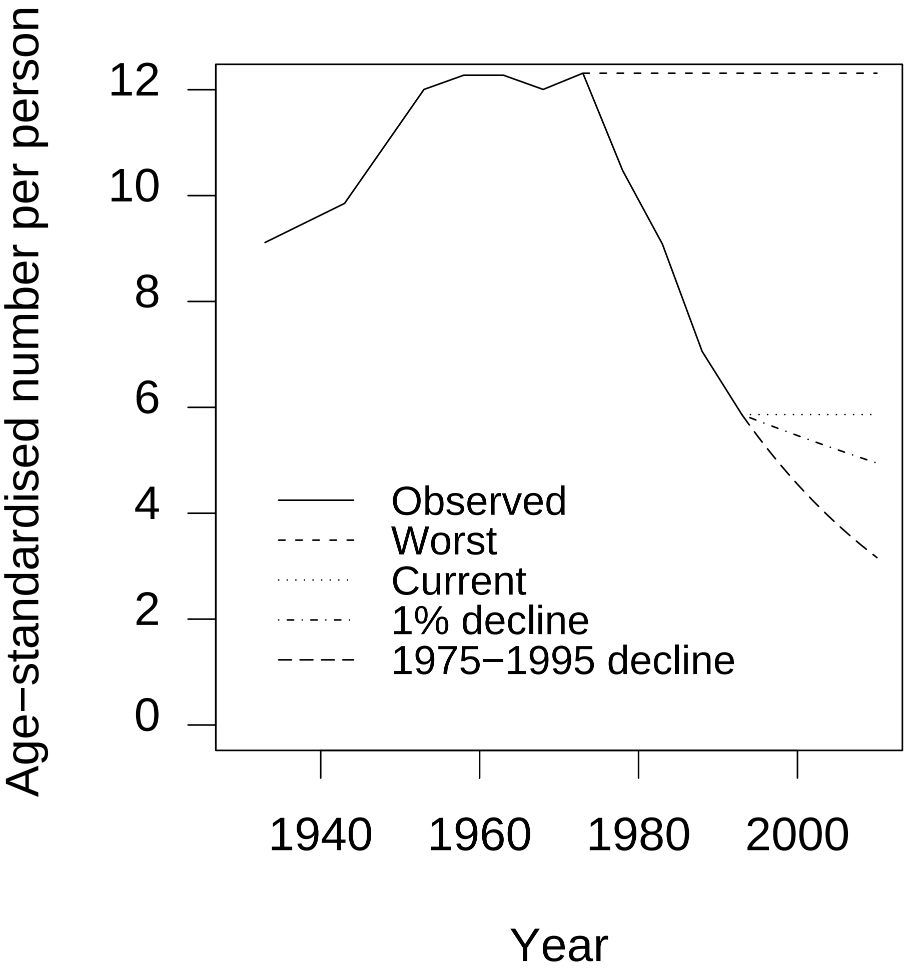


Figure A.5: Daily cigarette consumption for the different scenarios, males aged 30–74

Age-period modelling The most simple approach to forecasting is to project period rates of disease or mortality into the future using linear or curvi-linear functions [Hakama 1980; Hakulinen et al. 1986], and to assess the effects of using different functions or slopes which may correspond to likely different future scenarios of tobacco control. One simple approach has applied historical age-specific rates under different rate scenarios [Nam et al, 1996]. This method can encompass counterfactual scenarios in that past and future disease rates can be made to conform to observed maxima and minima, or plausible maxima and minima from published rates on other populations. The differences between the counterfactual hypotheses and the observed trends are the tobacco-attributable cases or deaths.

Age-cohort and age-period-cohort modelling Mortality from lung cancer during the twentieth century can be modelled to estimate the age-adjusted effects

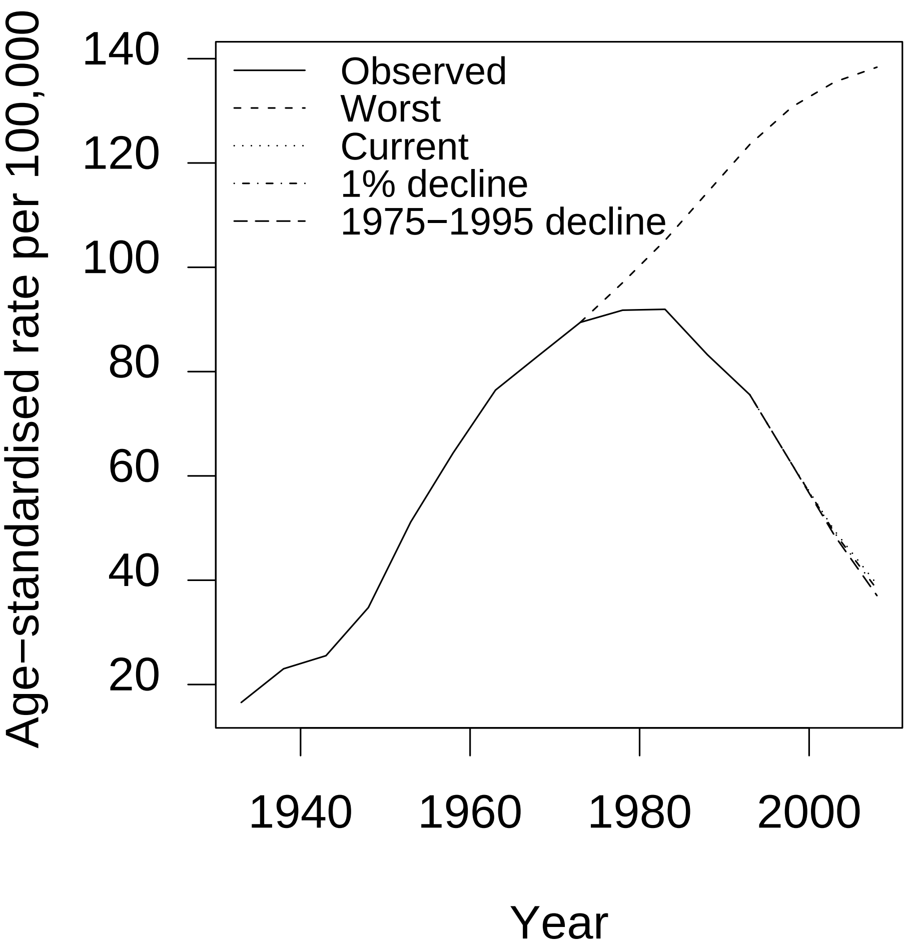


Figure A.6: Projected lung cancer mortality rates, males aged 30–74 years

of birth cohort (generation) and period (for each sex separately). Previous research suggests that birth cohort has a dominant influence in modelling lung cancer [Boyle, Roberstson 1987; Negri et al. 1990; Zheng et al. 1994; Lee et al. 1994; Reissigova et al. 1994; Lopez-Abente et al. 1995; Petrauskaite, Gurevicius 1996; Hristova et al. 1997; Jee et al. 1998], including in Australia [Taylor, McNeil 1997], and that this represents the differing cumulative exposures to causative factors of successive generations. It is assumed that the predominant cohort effect is tobacco smoking. Counterfactual scenarios would include estimation of past and future mortality rates if cohort effects did not decline, or continued to increase to certain levels observed in other populations with high exposures, or to levels which occur in cohorts of smokers. Alternatively, earlier or more intense tobacco control may have lead to earlier declines, lesser peaks and more rapid falls in cohort rates. The differences between the observed and the counterfactual hypotheses are the surmised effects of tobacco control.

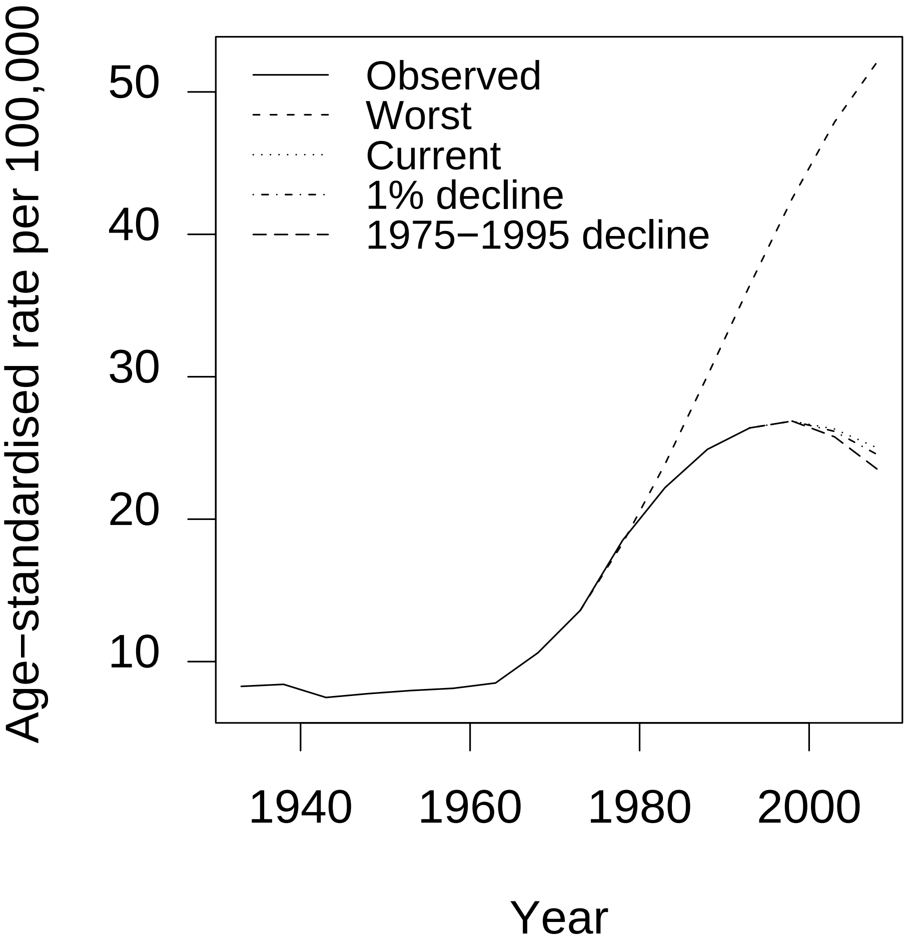


Figure A.7: Projected lung cancer mortality rates, females aged 30–74 years

(b) Estimation of tobacco effect by excess rates of lung cancer Rates of mortality from lung cancer by age group and sex are available for non-smokers from cohort studies. If these are subtracted from rates of lung cancer in the population then the rate of lung cancer attributable to smoking is estimated [Peto et al. 1992]. This method has been used to estimate tobacco-attributable mortality in a wide range of countries where tobacco smoking data are not available. Furthermore, it is claimed that the method may have advantages over use of tobacco smoking data since it actually assesses the results of cumulative tobacco exposure in populations [Peto et al. 1992]. Assuming the lung cancer mortality rate in non-smokers is constant over time, these calculations can be made for earlier periods to determine tobacco-attributable deaths in the population previously. Counterfactual scenarios involving differing peaks and declines of lung cancer mortality could be invoked to estimate tobacco-attributable deaths in different situations, including under different projections.

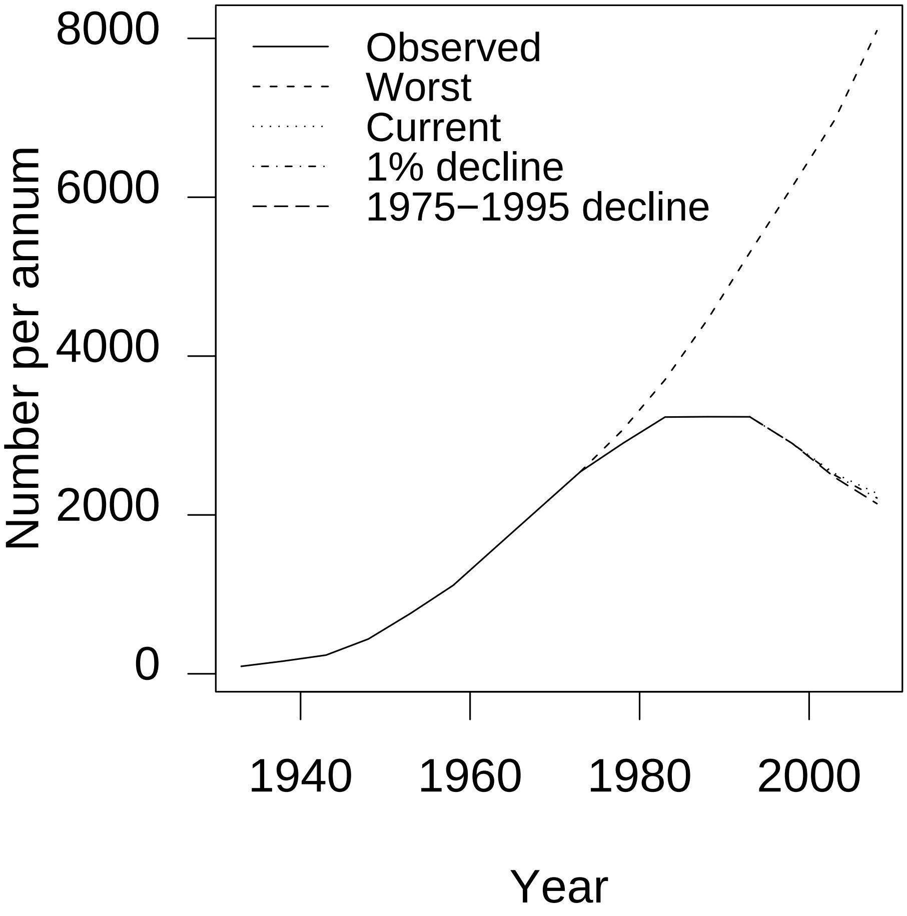


Figure A.8: Projected number of lung cancer deaths, males aged 30–74 years

Methods which require information on tobacco consumption

(a) Modelling tobacco consumption on lung cancer rates

Aggregate period modelling Many studies have examined the time trends (and differentials) in rates of lung cancer mortality in relation to aggregate tobacco consumption of populations, usually in a semi-quantitative way [Zaridze, Gurevicius 1986; La Vecchia et al. 1988; Kubik, Plesko 1998; Kublick et al. 1992].

Proxies for trends in tobacco consumption over long periods, such as data on cigarette or tobacco sales (which are often available because of tax or excise), can be expressed as per capita consumption (aged 15 years and over) — although for both sexes together. The delayed and cumulative effects of tobacco consumption are evident in the dis-junction (lag) in time between the curves of per capita tobacco consumption and those of lung cancer [Walker, Brin 1988]. These lags can be taken into account in regression analyses or time-series analyses which seek to relate overall

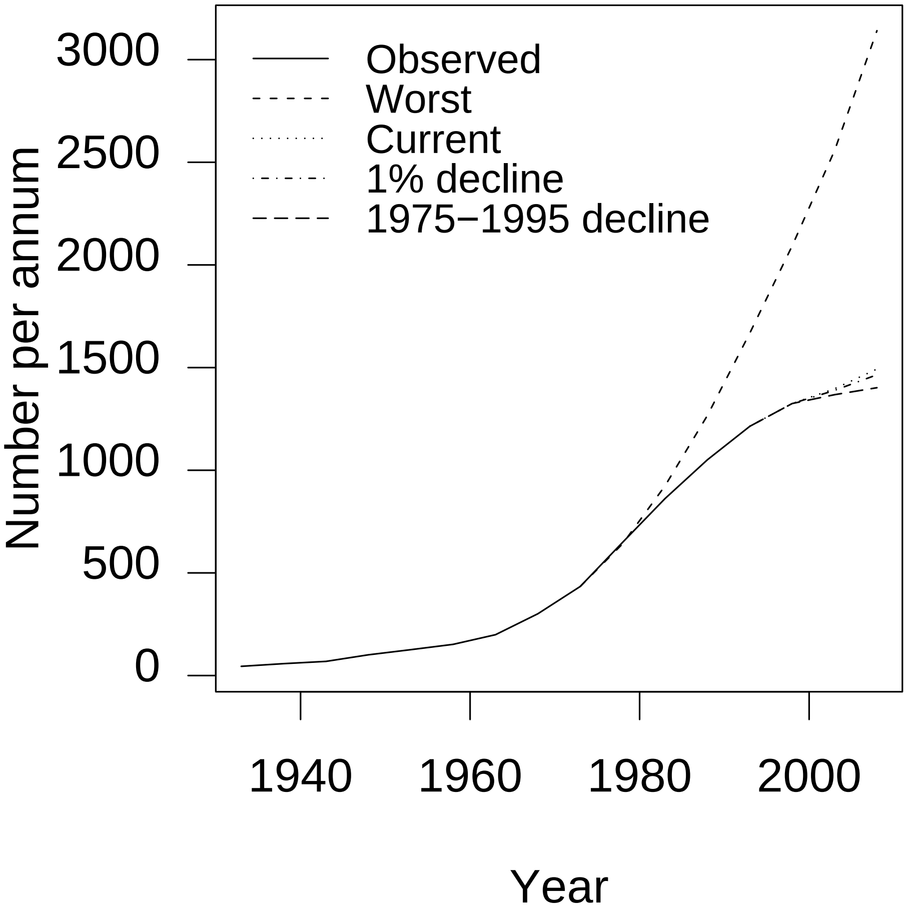


Figure A.9: Projected number of lung cancer deaths, females aged 30–74 years

tobacco consumption to lung cancer mortality [Hakama 1980, Hakama, Pukkala 1984], although this will not exclude the effects of causative exposure factors which co-vary with tobacco consumption. Pierce et al. [1992] predicted lung cancer rates in OECD countries based on the relationship between tobacco consumption and lung cancer two decades later.

If a satisfactory relationship can be established then counterfactual scenarios can be constructed which would enable the assessment of trends in mortality from lung cancer in Australia that would have occurred if tobacco consumption had not decreased, or if it had increased to levels observed in countries with high consumption.

Age-cohort / period-consumption modelling A more satisfactory method is to relate consumption of tobacco by successive birth cohorts to similar birth cohort rates of lung cancer mortality [Lee et al. 1990; Janssen-Heijnen et al. 1995; Nordlund 1998]. Brown and Kessler [1988] produced projections for USA using an age-period-consumption model, replacing the period effect with a period measure of

	Average annual deaths		Deaths averted	
	Actual	No decline in tobacco smoking	Annual	Cumulative
Males				
1971–1975	2543	2543	0	0
1976–1980	2900	3070	850	850
1981–1985	3232	3702	2350	3200
1986–1990	3236	4458	6110	9310
1991–1995	3235	5304	10345	19655
Females				
1971–1975	434	434	0	0
1976–1980	649	642	*	*
1981–1985	862	924	310	275
1986–1990	1051	1267	1080	1355
1991–1995	1214	1668	2270	3625

Table A.6: Lung cancer deaths averted from actual decline in tobacco smoking compared with maintenance of tobacco consumption at the observed maximum levels (* based on small numbers)

tar exposure. A similar approach was used by Negri et al. [1990] in Italy.

The most satisfactory approach is to relate cumulative tobacco consumption (by age) to age-specific rates of lung cancer by cohort [Doyle 1985]. Age structured regression models can be investigated which describe the association between cohort tobacco consumption and lung cancer mortality, and tobacco consumption can be varied to examine counterfactual scenarios in the past, and a range of plausible future situations with differing tobacco consumption. This is the method adopted in this report.

The Brown-Forbes model has been used for lung cancer predictions in the past [Mantel et al. 1986]. Predictions based on multistage models have been funded by the tobacco industry to show that predicted lung cancer mortality does not follow observed patterns of change [Swartz, 1992; Lee, Forey, 1998].

(b) Use of attributable fractions The quantitative attribution of lung cancer to tobacco smoking in populations is usually accomplished by the use of the attributable fraction (AF). These data suggest that the AF for male smoking is around 80% for lung cancer, for example. Combinations of these methods with population incidence or mortality data can be used to estimate absolute risk of tobacco-attributable disease [Taylor 1993a, 1993b].

Using estimates of smoking prevalence for different periods (or cohorts) it would be possible to estimate tobacco-attributable lung cancer deaths from available statistics, although account must be taken of changes of RR over time. For counterfactual scenarios, different AF could be calculated under different smoking prevalence assumptions, but these would have to be applied to mortality rates re-estimated from other methods which reflected such prevalences. Mao et al. [1992] used this method and calculated smoking-attributable deaths from 1969–2019 in Canada, with different prevalences of smoking, to indicate potential savings from tobacco control.

(c) Synthetic cohorts Yet another approach is to construct a mathematical model of a population over a century or more which incorporates tobacco smoking prevalences by age group, by sex, by period and cohort, and to estimate the ensuing rates of lung cancer deaths from data on the absolute risks of these conditions in relation to smoking intensity and duration derived from meta-analyses of case control and cohort studies [Hakulinen, Pukkala 1981; Manton et al. 1986; Swartz 1992]. This would need to take into consideration changes in RRs and absolute risk over time. The model would incorporate information on survival from other tobacco-related mortality and non-tobacco-related mortality. It would also be desirable to incorporate information on other exposures and their effects on lung cancer.

Assessment of estimation methods

Obviously methods which involve tobacco consumption data are superior to those which do not since the latter require assumptions that there are not significant other causes of lung cancer than tobacco smoking. Use of attributable fractions for these estimations is highly problematic and synthetic cohorts require many assumptions and do not rely on empirical data, except for calibration. We have selected an age-period model which also incorporates cumulative tobacco consumption by age for each birth cohort. This is superior to modelling period rates of lung cancer mortality on lagged period tobacco consumption.

A.4.2 Commentary on methods

A method has been presented whereby lung cancer mortality can be projected under different cigarette consumption scenarios. The worst case scenario for cigarette consumption would have seen lung cancer rates and numbers continue to climb appreciably.

The validity of the projections is dependent upon the validity of the risk model and of the scenario assumptions. First, tar adjustment was not undertaken. Evidence suggests that after adjustment for number of cigarettes smoked, rate ratios for smoking high tar cigarettes are not significant [Wilcox et al. 1988]. Moreover, smokers of low tar cigarettes may compensate by smoking more cigarettes or smoking more deeply [Wilcox et al. 1988]. There is also evidence that the design of cigarettes has changed so that tar yields from standard tests may not provide an accurate estimates of tar availability [Wilkenfeld et al 2000]. Second, the historical patterns of smoking for females was based on a hypothesised pattern. Although the pattern was based on available information, the information was limited and the levels for the 1930s and 1940s may not be accurate. Third, the adjustment for under-diagnosis was based on uncertain data. Fourth, it is unclear whether the assumptions of the age and period effects being stable out into time are valid, but in this analysis projections are only made to 2010.

A.4.3 Commentary on results

The projections point to broad patterns of change in lung cancer mortality. Male lung cancer mortality rates can be expected to continue to decline, while female rates can be expected to peak and then decline. These changes are largely an expression of successful tobacco control efforts over the past 30 years. The worst scenario suggests that an absence of tobacco control measures would have had an even more devastating impact on the health of Australians comparable to rate reached in the UK [Coleman et al. 1993].

Under the different scenarios, lung cancer mortality could reach quite different levels in the future. Therefore tobacco control measures today can have a profound impact on lung cancer tomorrow. Vigilance is required to ensure that cigarette consumption rates do not begin to rise again. Any rise in tobacco consumption will not be expressed by changes in lung cancer rates for 20–30 years.

A.5 References

ABS. *Population Projections, 1997 to 2051.* ABS Catalogue 3222.0. Canberra: Australian Bureau of Statistics, 1998.

Boyle P, Robertson C. Statistical modelling of lung cancer and laryngeal cancer incidence in Scotland, 1960–1979. *Am J Epidemiol*, 125(4):731–44, 1987.

Brown CC, Kessler LG. Projections of lung cancer mortality in the United States: 1985–2025. *J Natl Cancer Inst*, 80(1):43–51, 1988.

Brown K, Forbes F. A mathematical model of aging processes. *Gerontology*, 29:46–51, 1974.

Coleman M, Esteve J, Damiecki P, et al. *Trends in Cancer Incidence and Mortality*. Lyon: International Agency for Research on Cancer, 1993.

Doyle EJ. *A Cohort Analysis of Smoking and Lung Cancer in Australia, Canada and the United Kingdom*. University of Newcastle, NSW, Australia. 1985. PhD Thesis.

Dwyer T, Blizzard L, Shugg D, Hill D, Ansari MZ. Higher lung cancer rates in young women than young men: Tasmania, 1983 to 1992. *Cancer Causes & Control*, 5(4):351–8, 1994.

English D, Holman D, Milne E, et al. *The Quantification of Drug Caused Morbidity and Mortality in Australia, 1995 edition*. Canberra: Commonwealth Department of Human Services and Health, 1995.

Giles GG, Hill DJ, Silver B. The lung cancer epidemic in Australia, 1910 to 1989. *Aust J Public Health*, 15(3):245–7, 1991.

Gray N, Hill D. Patterns of tobacco smoking in Australia. 2. *Med J Aust*, 2 (10):327–328, 1977.

Gray N, Hill D. Patterns of tobacco smoking in Australia. *Med J Aust*, 2(22):819–822, 1975.

Hakama M, Pukkala E. The projection of chronic diseases using data on risk factors and risk factors intervention: the case of cancer. *World Health Statistical Quarterly*, 37(3):318–327, 1984.

Hakama M. Projection of cancer incidence: experiences and some results in Finland. *World Health Statistical Quarterly*, 33(4):228–40, 1980.

Hakulinen T, Pukkala E. Future incidence of lung cancer: forecasts based on hypothetical changes in the smoking habits of males. *Int J Epidemiol*, 10:233–240, 1981.

Hakulinen T, Teppo L, Saxén E. Do the predictions of cancer incidence come true? Experience from Finland. *Cancer*, 57:2454–2458, 1986.

Hill D, Gray N. Australian patterns of tobacco smoking and related health beliefs in 1983. *Community Health Stud*, 8(3):307–316, 1984.

Hill D, Gray N. Patterns of tobacco smoking in Australia. *Med J Aust*, 1(1):23–25, 1982.

Hill D, White V, Gray N. Australian patterns of tobacco smoking in 1989. *Med J Aust*, 154(12):797–801, 1991.

Hill D, White V, Gray N. Measures of tobacco smoking in Australia 1974–1986 by means of a standard method. *Med J Aust*, 149(1):10–12, 1988.

Hill D, White V, Scollo M. Smoking behaviours of Australian adults in 1995: trends and concerns. *Med J Aust*, 168(5):209–213, 1998.

Hill D. Australian patterns of tobacco smoking in 1986. *Med J Aust*, 149(1):6–10, 1988.

Holman C, Armstrong B. *Cancer Mortality Trends in Australia 1910–1979*, Perth: Cancer Council of Western Australia, 1982.

Holman C, Hatton W, Armstrong B, English D. *Cancer Mortality Trends in Australia Volume II 1910–1984*, Perth: Health Department of Western Australia, 1984.

Hristova L, Dimova I, Iltcheva M. Projected cancer incidence rates in Bulgaria, 1968–2017. *Int J Epidemiol*, 26(3):469–75, 1997.

Ireland AW, Lawson JS. The changing face of death: recent trends in Australian mortality. *Med J Aust*, 1(12):587–90, 1980.

Janssen-Heijnen ML, Nab HW, van Reek J, van der Heijden LH, Schipper R, Coebergh JW. Striking changes in smoking behaviour and lung cancer incidence by histological type in south-east Netherlands, 1960–1991. *European J Cancer*, 31A(6):949–52, 1995.

Jee SH, Kim IS, Suh I, Shin D, Appel LJ. Projected mortality from lung cancer in South Korea, 1980–2004. *Int J Epidemiol*, 27(3):365–9, 1998.

Kubik A, Hakulinen T, Reissigova J, Luostarinen T. Lung cancer and smoking in Finland and the Czech Republic. *Neoplasma*, 39(3):177–84, 1992.

Kubik A, Plesko I. Trends in cigarette sales and lung cancer mortality in four central European countries. *Central European J Public Health*, 6(1):37–41, 1998.

La Vecchia C, Levi F, Decarli A, Wietlisbach V, Negri E, Gutzwiller F. Trends in smoking and lung cancer mortality in Switzerland. *Preventive Medicine*, 17(6):712–24, 1988.

Lee LT, Lee WC, Lin RS, Chen SC, Chen CY, Luh KT, Chen CJ. Age-period-cohort analysis of lung cancer mortality in Taiwan, 1966–1990. *Anticancer Research*, 14(2B):673–6, 1994.

Lee P, Forey P. Trends in cigarette consumption cannot fully explain trends in British lung cancer rates. *J Epidemiol Community Health*, 52 (2):82–92, 1998.

Lee PN, Fry JS, Forey BA. Trends in lung cancer, chronic obstructive lung disease, and emphysema death rates for England and Wales 1941–85 and their relation to cigarette smoking. *Thorax*, 45(9):657–665. 1990

Lopez-Abente G, Pollan M, de la Iglesia P, Ruiz M. Characterisation of the lung cancer epidemic in the European Union (1970–1990). *Cancer Epidemiology, Biomarkers & Prevention*, 4(8):813–20, 1995.

Mantel H, Forbes W, Thompson M, Gibberd R. Quantitative models of lung cancer mortality. II. Predicting lung cancer mortalities for a population depending on the level of smoking. *Can. J Public Health*, 77 (3):208–215, 1986.

Manton KG, Stallard E, Creason JP, Riggan WB, Woodbury MA. Compartment model approaches for estimating the parameters of a chronic disease process under changing risk factor exposures. *Comput Biomed Res*, 19(2):151–69 1986.

Mao Y, Gibbons L, Wong T. The impact of the decreased prevalence of smoking in Canada. Canadian Journal of Public Health. *Revue Canadienne de Sante Publique*, 83(6):413–6, 1992.

Nam C, Roger R, Hummer R. Impact of future cigarette smoking scenarios on mortality of the adult population in the United States, 2000–2050. *Soc Biol*, 43:155–168, 1996.

NCI Monograph No. 8. *Changes in cigarette-related disease risks and their implication for prevention and control.* Bethesda (MD): U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute; 1997 (NIH Publ No. 97–4213).

Negri E, La Vecchia C, Decarli A, Boyle P. Projections to the end of the century of mortality from major cancer sites in Italy. *Tumori*, 76(5):420–8, 1990.

Nordlund LA. Trends in smoking habits and lung cancer in Sweden. *European Journal of Cancer Prevention*, 7(2):109–16, 1998.

Peto R, Lopez A, Boreham J, Thun M, Heath C (Jnr). Mortality from tobacco in developed countries : indirect estimation from national vital statistics. *Lancet*, 339: 1268–1278, 1992.

Petrauskaite R, Gurevicius R. Time trends in lung-cancer mortality rates among men in Lithuania, 1965–1994. *Int J Cancer*, 66(3):294–6, 1996.

Pierce JP, Thurmond L, Rosbrook B. Projecting international lung cancer mortality rates: first approximations with tobacco-consumption data. *Journal of the National Cancer Institute*, Monographs, (12):45–9, 1992.

Reissigova J, Luostarinen T, Hakulinen T, Kubik A. Statistical modelling and prediction of lung cancer mortality in the Czech and Slovak Republics, 1960–1999. *Int J Epidemiol*, 23(4):665–72, 1994.

Rohan T, Christie D. Australian cancer mortality from 1950 to 1977: analysis of the national mortality statistics. *Med J Aust*, 1(3):109–13, 1980.

SAS Institute Inc. *SAS/STAT Software: Changes and Enhancements through Release 6.11*, Cary, NC: SAS Institute Inc., 1996.

Stanley K, Stjernsward J, Koroltchouk V. Cancers of the stomach, lung and breast: mortality trends and control strategies. *World Health Statistics Quarterly — Rapport Trimestriel de Statistiques Sanitaires Mondiales*, 41(3–4):107–14, 1988.

Stevens R, Moolgavkar S. A cohort analysis of lung cancer and smoking in British males. *Am J Epidemiol*, 119:624–641, 1984.

Swartz J. Use of a multistage model to predict time trends in smoking induced lung cancer. *J Epidemiol Community Health*, 46 (3):311–315, 1992.

Taylor R, McNeil D. *Projections of incidence of major cancers in NSW to 2010.* NSW Cancer Council. May 1997

Taylor R. (1993a) Estimation of tobacco-induced mortality from readily available information. *Tobacco Control*, 2:18–23, 1993.

Taylor R. (1993b) Risks from premature deaths from smoking in 15 year old Australians. *Aust J Public Health*, 17(4):358–364. 1993.

Thom TJ, Epstein FH. Heart disease, cancer, and stroke mortality trends and their interrelations. An international perspective. *Circulation*, 90(1):574–82, 1994.

Todd G. Cigarette consumption per adult of each sex in various countries. *J Epidemiol Community Health*, 32 (4):289–293, 1978.

Vakil DV. Lung cancer mortality trends in Canada from 1931 to 1982. *Cancer Detection & Prevention*, 13(2):87–93, 1988.

Vutuc C, Gredler B. Lung cancer in Austria: present and future trends. *European J Epidemiol*, 2(2):158–162. 1986.

Walker WJ, Brin BN. U.S. lung cancer mortality and declining cigarette tobacco consumption. *J Clin Epidemiol*, 41(2):179–85, 1988.

Wilcox H, Schoenberg J, Mason T, Bill J, Stemhagen A. Smoking and lung cancer: risk as a function of cigarette tar content. *Prev Med*, 17:263–272, 1988.

Wilkenfeld J, Henningfield J, Slade J, Burns D, Pinney J. It's time for a change: cigarette smokers deserve meaningful information about their cigarettes. *J Natl Cancer Inst*, 92 (2):90–92, 2000.

Winstanley M, Woodward S, Walker N. *Tobacco in Australia, Facts and Issues, 1995*. Australia: Quit Victoria, 1995.

Zaridze DG, Gurevicius R. *Lung cancer in the USSR: patterns and trends.* IARC Scientific Publications (Lyon). (74):87–101, 1986.

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*, 74(5):1556–67, 1994.

Appendix B

Appendix: Will US smoking prevalence “inexorably continue to decline”?

Abstract

Aim: To assess whether “the demographics of smoking imply that [US] prevalence will inexorably continue to decline over the next several decades” (Mendez et al, Am J Epidemiol 1998;148:249–258), we re-analysed the same data using different techniques to check the authors’ assumptions. **Method:** The rate of smoking cessation was split between changes in prevalence by age and by time, and differential mortality. Prevalence was modelled using generalised linear models, with changes by age assessed using the 1997 National Health Interview Survey (NHIS) and temporal changes using 1983–1995 NHIS. The effect of ageing was shown by weighting age-specific prevalence by population projections. **Results:** Mendez et al and this study obtained different estimates for temporal changes in prevalence from 1983–1993. We found that temporal changes were less during 1991–1993 than 1983–1989. Population ageing had little impact on adult smoking prevalence (2–4% over 50 years). **Discussion:** The validity of estimates for rates of cessation was questioned, where better data are required. Mendez et al may have over-estimated temporal changes in prevalence. Demographics of smoking do not necessarily imply that smoking prevalence will inexorably continue to decline.

B.1 Introduction

Tobacco smoking is an important determinant of health. Prediction of smoking behaviour for the future allows for setting targets for smoking programmes for today and assessing the health impact for tomorrow.

A recent editorial in the American Journal of Public Health (Green et al., 2000) [1] used results from Mendez et al. (1998) and Mendez and Warner (2000) to show it was implausible that US national smoking prevalence targets would be achieved. Mendez et al. (1998) had earlier concluded that "the demographics of smoking imply that [US] prevalence will inexorably continue to decline over the next several decades." This confident conclusion was based on the development of a dynamic smoking model, including effects due to differential mortality, smoking initiation and smoking cessation (see Figure B.1). The cessation parameters were estimated by finding the non-linear least squares fit for the dynamic model.

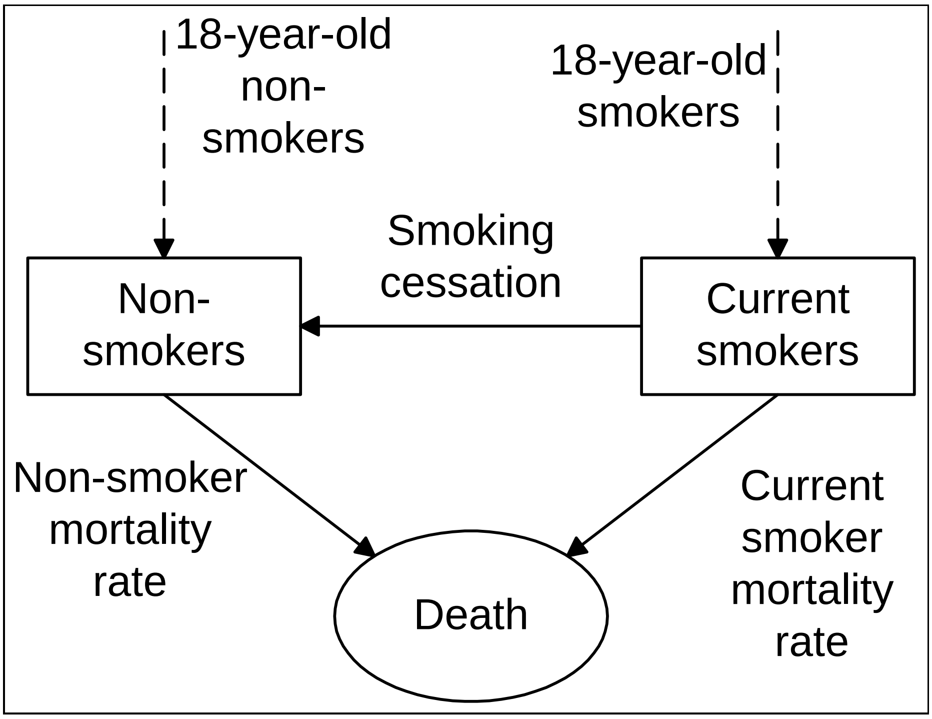


Figure B.1: Dynamic smoking model, Mendez et al. (1998)

The authors assumed that the average cessation rate during 1983–1993 would be representative for later years. This was based on a test for change in trend between 1983–1988 and 1990–1993, which found no evidence for change. However, more recent data from the CDC (1999) suggest that the decline in adult prevalence and

[1] Citations in this appendix refer to the main bibliography.

age-specific prevalence may have lessened or possibly ceased. Similar patterns have been observed elsewhere, including Australia (Hill et al., 1998) and New Zealand (Ministry of Health, 1999a).

An alternative method is presented whereby the net rate of cessation is decomposed into additive components for the decline in age-specific prevalence with respect to period, the decline in period-specific prevalence with respect to age, and the mortality differential between current smokers and the total population. This alternative approach will allow some assessment of the validity of the estimates used by Mendez and colleagues. Similar data sources were used to Mendez et al, with supplemental data from later National Health Interview Surveys and rate ratios for ex-smokers.

A simple assessment of the effect of ageing of the population on total adult prevalence using population projections is also presented.

B.2 Methods

The data include: prevalence results from the 1970–1995 and the 1997 National Health Interview Surveys (Mendez et al., 1998; National Center for Health Statistics, 2000); life tables for the US for 1997 (Anderson, 1999); rate ratios for current smoking from the 1986 National Mortality Followback Study (Rogers and Powell-Griner, 1991); rate ratios for ex-smokers from a meta-analysis (Holman et al., 1990); 1997 population estimates by age (U.S. Census Bureau, 2000b) and 2050 population projections by age (U.S. Census Bureau, 2000a).

The analytical approach used the relationship for a cohort that:

$$\begin{aligned}\text{rate of cessation} &= \text{partial rate of decline of prevalence by time}\\ &+ \text{partial rate of decline of prevalence by age}\\ &- \big(\text{ current smoker mortality rate}\\ &\quad - \text{total mortality rate }\big) \end{aligned} \tag{B.1}$$

which is derived in Chapter 3. The first component is time-dependent, while the latter three components were assumed constant over time and are described here as the *static rate of cessation.* In outline, the rate of decline of prevalence by time was estimated using data from the National Health Interview Surveys for 1983–1995. The rate of decline of prevalence by age was estimated from the 1997 National Health Interview Survey. Estimates for differential mortality used prevalence for current

	Age group (years)				
Year	18–24	25-44	45-64	≥65	Overall
1994	0.275	0.300	0.255	0.120	0.255
1995	0.248	0.286	0.255	0.130	0.247

Table B.1: Prevalence of current smoking among US adults from National Health Interview Survey (1994, 1995)

smokers (π_c) and ex-smokers (π_x) from the 1997 National Health Interview Survey, mortality rates from vitals and rate ratios from the literature.

To investigate the rate of change of age-specific smoking prevalence over time, data were taken from the National Health Interview Surveys for 1983–1995 (see Mendez et al., 1998) and Table B.1). The age groups were 18–24 years, 25–44 years, 45–64 years, 65 years and over. The smoking question was changed in 1992 and prevalence results have been adjusted by subtracting one percent from surveys since that time (Mendez et al., 1998). As the 1997 survey design was substantially revised, those data were excluded.

Effect sample sizes for each year and age group were estimated by taking conservative variances from Mendez et al. (1998) together with the prevalence of current smoking:

$$\text{Effective cell size} = \frac{\pi_c(1-\pi_c)}{\text{variance}}.$$

Log-binomial regression (Skov et al., 1998) was used to ascertain whether there was a significant change in slope between 1983–1990 and 1990–1993 or 1990-1995 using piecewise generalized linear regression (Chu et al., 1999; SAS Institute Inc., 1999). A test was also made as to whether there was a significant change in prevalence across 1990–1993 or 1990–1995, and whether the change for the 1990s was different between age groups. Parameters from the regression estimate the rate of decline of age-specific prevalence of current smoking by age group for the periods 1983–1993, 1990–1993 and 1990–1995.

Prevalence results of former and current smoking by single years from the 1997 NHIS were modelled using multinomial regression separately by sex, taking account of the complex survey design. This was done using `proc multilog` in SUDAAN (Shah et al., 1996). The fitted parameter estimates were used to calculate the predicted prevalence of current smoking, prevalence of ex-smoking and the rate of decline of period-specific prevalence by age.

To estimate the mortality differential between current smokers and the total population, estimates of all-cause mortality (μ) by single year of age and by sex were taken from life tables. The mortality rate for never smokers (μ_0) was estimated from prevalence estimates and from the rate ratios of current smokers (RR_c) and ex-smokers (RR_x) by

$$\mu_0 = \frac{\mu}{1 + \pi_c(RR_c - 1) + \pi_x(RR_x - 1)}.$$

The values for RR_c and RR_x by single years of age were estimated by loess smoothing (Cleveland and Devlin, 1988) using `R` software (Ihaka and Gentleman, 1996). The age and sex specific rate of cessation was then calculated using the relationship in Equation (B.1).

Variance estimates for the static rate of cessation were calculated using the delta method (Rao, 1973). It was assumed that the mortality rate, the coefficients from the multinomial regression and the rate ratios were independent of each other. The covariance matrix for the delta method was therefore block diagonal, using the covariance matrix from the multinomial regression, the variance for the mortality rate assuming a Poisson distribution, and zero values for the variance of the rate ratios. The partial differential matrix was estimated numerically using `R` software. Details are given in the following section.

Variance estimates for the rate of cessation assumed that static rate of cessation and the rate of decline in prevalence by time were independent. Point estimates and variances for an age group were estimated by weighting the results for each single year of age by the estimated number of current smokers.

Total adult smoking prevalence from the National Health Interview Surveys for 1970–1997 was modelled over time for the sake of illustration using log-binomial models and local logistic regression models using the `locfit` package (Loader, 1999) in `R` software.

The effect of the ageing population was shown by weighting the age-specific prevalence for 1970 and 1997 by the middle population projections for 2050.

B.2.1 Variance estimation

We want to find point estimates and covariances for the static cessation rate at different ages. The rate ratios RR_c and RR_x have been assumed to be stochastically independent, as the meta-analytic approach taken does not allow for an evaluation of this assumption. It has further been assumed that the mortality rate μ follows a Poisson distribution estimate from a population with d deaths, which have been collected with limited systematic variation.

$$\mathrm{var}(\mu) = \frac{\mu^2}{d}$$

One methodological issue is that π_c and π_x have a multinomial distribution. As there is no inherent ordering of current, ex- and never smokers, an appropriate model for these nominal data would be multinomial logistic regression using generalised logits. For a group defined by covariate vector x, the prevalence for being a current or former smoker, respectively, will be

$$\begin{aligned}\pi_c &= \frac{\exp(\boldsymbol{\beta}_c'\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}_c'\boldsymbol{x})+\exp(\boldsymbol{\beta}_x'\boldsymbol{x})},\\ \pi_x &= \frac{\exp(\boldsymbol{\beta}_x'\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}_c'\boldsymbol{x})+\exp(\boldsymbol{\beta}_x'\boldsymbol{x})}\end{aligned}$$

and the prevalence for never smokers is $1-\pi_c-\pi_x$. From the parameters from the multinomial regression, it is possible to estimate the rate of decline of prevalence of current smoking, where

$$-\frac{\partial\pi_c}{\partial a}/\pi_c = \left[\pi_x\boldsymbol{\beta}_x-(1-\pi_c)\boldsymbol{\beta}_c\right]'\frac{\partial\boldsymbol{x}}{\partial a}.$$

Given these values, together with μ, RR_c and RR_x, the point estimate for $-\frac{\partial\pi_c}{\partial a}/\pi_c-(RR_c\mu_0-\mu)$ can then be calculated.

Estimates of the covariance matrix for the predicted values were estimated using the delta method (Rao, 1973). Specifically, let the set of parameters be

$$\boldsymbol{\theta} = \begin{pmatrix}\boldsymbol{\beta}_c\\ \boldsymbol{\beta}_x\\ \mu\\ RR_c\\ RR_x\end{pmatrix}$$

and let the point estimate as a function of the parameters be

$$\boldsymbol{g}(\boldsymbol{\theta}) = -\frac{\partial\pi_c}{\partial a}/\pi_c-(RR_c\mu_0-\mu).$$

Note that $\boldsymbol{g}$ here is only a scalar, however the vector form used by Rao (1973) has been kept for generality. For notational simplicity, let

			$H_0 : \beta_1 = \beta_2$	
Piecewise periods	β_1	β_2	χ^2	p-value
1983–1990, 1990–1993	0.0304	0.0197	4.39	0.0361
1983–1990, 1990–1995	0.0307	0.0169	17.19	< 0.0001

Table B.2: Tests for changes in rates of decline in prevalence by time between the first period (β_1) and the second period (β_2), adjusted for age

$$\boldsymbol{\Sigma}_\beta = \text{var}\begin{pmatrix} \boldsymbol{\beta}_c \\ \boldsymbol{\beta}_c \end{pmatrix}$$

so that the covariance matrix for the parameter vector $\boldsymbol{\theta}$ is

$$\boldsymbol{\Sigma} = \text{var}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\Sigma}_\beta & 0 & 0 & 0 \\ 0 & \frac{\mu^2}{d} & 0 & 0 \\ 0 & 0 & \sigma^2_{RR_c} & 0 \\ 0 & 0 & 0 & \sigma^2_{RR_x} \end{pmatrix}.$$

By introducing the matrix of partial derivatives

$$\boldsymbol{G} = \left(\frac{\partial g_i}{\partial \theta_j}\right),$$

we can estimate the desired variance

$$\text{var}(\boldsymbol{g}) = \boldsymbol{G\Sigma G}'.$$

B.3 Results

To illustrate the decomposition of change of prevalence by age and by time, some of the years are shown in Figure B.2. This illustrates that the data before 1997 were highly aggregated and “lumpy”. The change in age-specific prevalence by time is a vertical movement, while the change in period-specific prevalence by age is a movement along the 1997 age-specific curve. Qualitatively, the changes in the aggregated data over time are crude, while the changes by age in the 1997 data are comparatively more precise.

There is good evidence that the temporal rate of decline of prevalence during the 1983–1990 period is different from the rate of decline during the 1990s (Table B.2). This is at variance with results from Mendez et al. (1998).

Tests for age interactions were consistent that the period rate of decline for those aged 18–24 years was different from the rate for those aged 25 years and over. Period

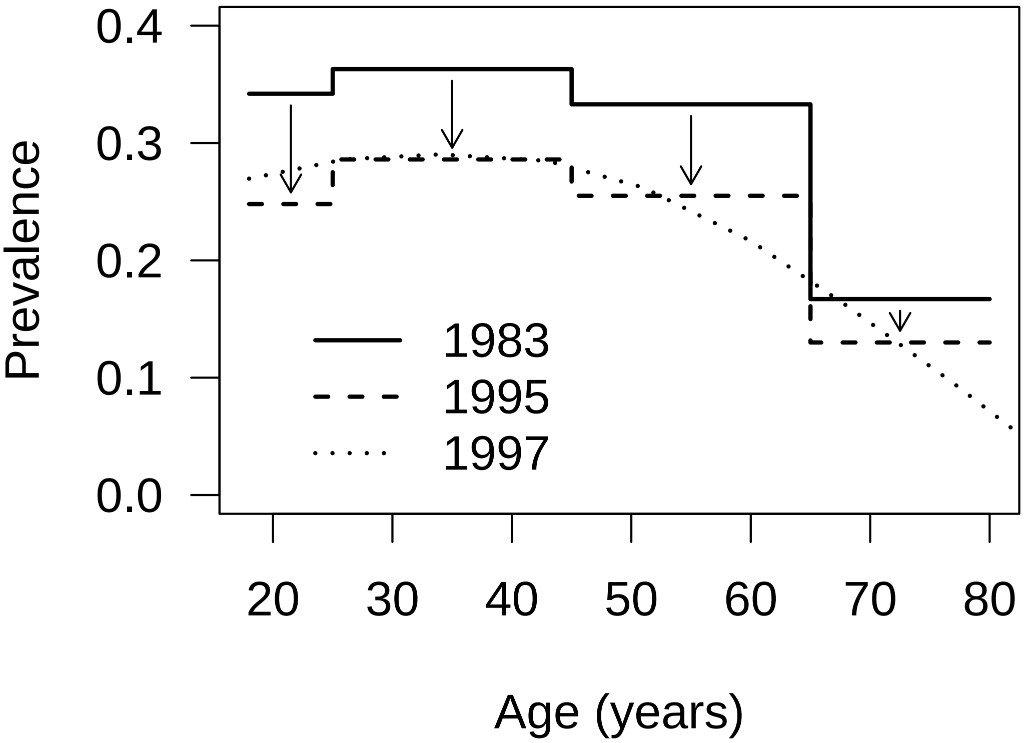


Figure B.2: Modelling US smoking prevalence by changes in age and time

			95% confidence interval	
Period	Age (years)	Estimate	Lower limit	Upper limit
1983–1993	18–24	0.0339	0.0293	0.0386
	25–84	0.0266	0.0245	0.0288
1990–1995	18–24	-0.0079	-0.0196	0.0038
	25–84	0.0197	0.0143	0.0251

Table B.3: Period rates of decline of age-specific prevalence for US adults, by age group

rates and confidence intervals have been estimated for the periods 1983–1993 and for 1990–1995 (Table B.3). Importantly, the rates of decline were slower for the more recent period (1990–1995), and there was even an increase in prevalence for those aged 18–24 years. The estimated period rates changed by 0.047 for those aged 18–24 years and by 0.013 for those aged 25 years and over between the two analysis periods.

There is little definition of the rates of decline for different age groups, so that the average rate of decline for a broad age group is precise, however the rate of decline for a specific age may not be valid.

Estimates of the rates of cessation are shown in Table B.4, together with the comparable results from Mendez et al. (1998).

			95% confidence interval	
Period	Age (years)	Estimate	Lower limit	Upper limit
1981–1993	18–30	0.027	0.025	0.030
	31–50	0.029	0.028	0.030
	51–84	0.053	0.052	0.055
1990–1995	18–30	0.002	0.000	0.005
	31–50	0.022	0.021	0.023
	51–84	0.046	0.045	0.048
Mendez et al. (1998)	18–30	0.002	-0.007	0.011
	31–50	0.021	0.016	0.027
	≥ 51	0.06	0.054	0.065

Table B.4: Comparison of cessation rates for US adults

Taking a simple approach by modelling for total prevalence, similar results to Mendez and colleagues were obtained for the period 1970–1993 (see Figure B.3). In particular, the local regression smoothing follows the earlier results by Mendez et al well until the 1990's. It is unclear whether the 1997 estimate is comparable to earlier years. The short-term projection using the slope for 1983–1993, approximating the projection for Mendez et al varies from the short-term projections using the slopes for 1990–1993 and for 1990–1995.

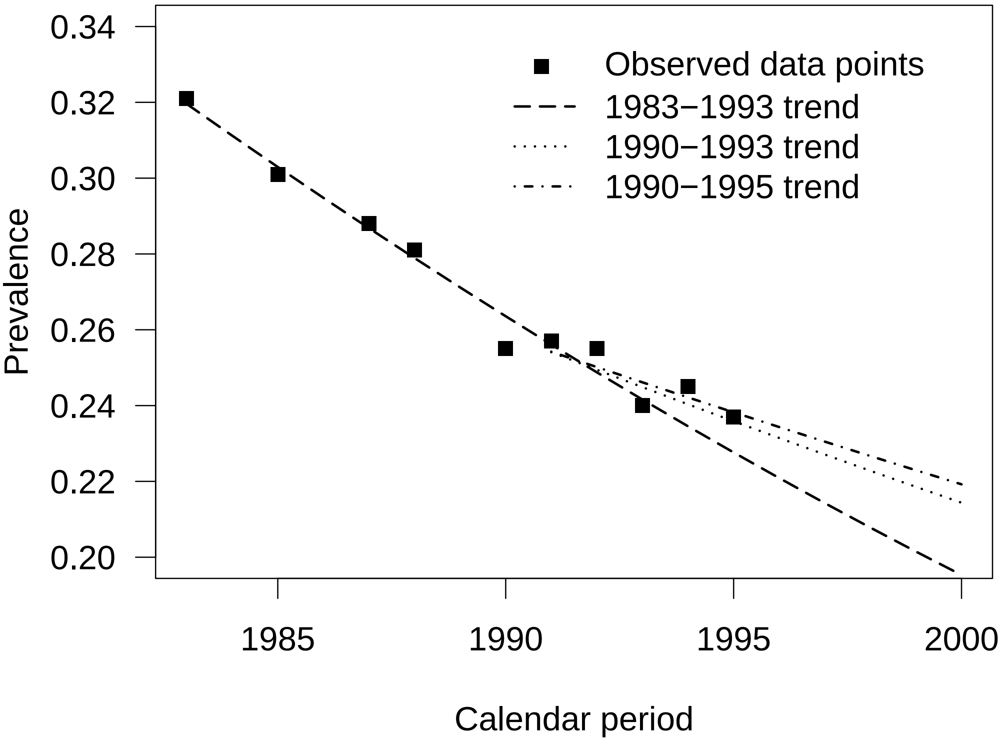


Figure B.3: US smoking prevalence fitted with different models

If 1997 age-specific prevalence were to be stable, the ageing of the population would drive adult prevalence from 24.7% in 1997 down to approximately 23.0% in 2050. Similarly, the 1970 adult smoking prevalence of 37.4% would have reduced to approximately 34.5% by 2050 were the same age-specific prevalence pattern to be observed.

B.4 Discussion

In summary, there is good evidence that rates of decline of prevalence, and hence the rates of cessation, have changed between the 1980's and the 1990's. The rates of cessation for US adults estimated by Mendez and colleagues may therefore be biased. Moreover, the effect of ageing of the population alone would only reduce prevalence by a few percentage points.

Before discussing possible implications, we consider potential issues with the study design. First, the smoking prevalence data from the National Health Interview Survey were only available by broad age groups, as can be seen graphically from Figure B.2. The age groups therefore only provide a summary of changes by age. Moreover, there is limited ability to find interactions between changes over time by age. One possible solution to this problem would be to obtain prevalence data by five year age groups, as has been used recently for cohort prevalence estimation (National Cancer Institute, 1997).

The main advantage of the current method is that the data were modelled taking account of the age-aggregated nature of the data, with supplemental age data from the 1997 National Health Interview Survey.

Second, the analysis assumed that the rate ratios were valid and precise. There is a surprising range of estimates for rate ratios from the literature (Holman et al., 1990; National Cancer Institute, 1997; English et al., 1995). 1986 National Mortality Followback Study rate ratio estimates for current smokers have been taken as the most representative in the US population in recent years, while the estimates for ex-smokers appear high compared with the literature.

Third, Mendez et al. (1998) correctly point out that the estimated measure is *net* cessation. Results for younger males and females suggest that they continue to take up smoking, so that the net cessation rate may not be a sensible measure for these ages. An alternative approach would be to restrict analysis to older ages where uptake is less.

These first three issues apply equally to this study and the study by Mendez and colleagues.

Fourth, the decline in cohort prevalence was estimated by separating out changes by time and changes by age. The main assumption of this study is that any age and period interaction is constant over time. This assumption is difficult to verify, although any change in an interaction is expected to be small.

The two studies derived estimates of similar precision for broad age groups. Due to data quality, the cessation rates modelled by Mendez et al and the change in age-specific prevalence by age in this study were by necessity means for aggregate age groups, which is unlikely to be realistic.

The two studies derived different estimates for the same period (1981–1993). This may be surprising, because the underlying dynamic model and data sources were similar between the two studies. Irrespectively, using the same method, different estimates were obtained from different periods in this study. This suggests that changes in rates of cessation may be rapid and that detailed estimates of cessation are difficult to obtain.

The two main issues with the study by Mendez and colleagues are data quality, as already discussed, and the interpretation of the results. Given differences in results, there is insufficient evidence to be confident that smoking cessation rates are stable, which is the main assumption for the projections made by Mendez et al.

There have previously been few attempts to quantify the rate of cessation in the mainstream epidemiological literature. Recent efforts using NHIS data to estimate cessation rates found considerable variability in rates between cohorts over time (National Cancer Institute, 1997).

Most previous efforts to estimate prevalence have used a variety of cohort-based models. The decomposition in Equation (B.1) supports the dominance of this simple approach. In short, implicit assumptions about static mortality patterns and static age by period smoking interactions suggest changes in adult smoking prevalence describe changes in cessation. Taking this further, modelling age-specific or age-standardised smoking prevalence will offer similar estimates to Mendez and colleagues (see Figure B.3).

Although the model formulation and estimation by Mendez and colleagues was elegant, poor data constrain the interpretation of their results. As suggested recently by Gunning-Schepers (1999), more complex modelling of health dynamics behoves the investigator to take greater care in understanding the complexity of systems. An incorrect mathematical representation or inaccurate data may lead to misleading conclusions.

These results suggest that the cessation rates estimated by Mendez et al can be interpreted as being optimistic. One implication is that the qualitative conclusions reached recently by two of the authors (Mendez and Warner, 2000) will continue to be true: the national smoking targets rates are unlikely to be achieved based on optimistic rates for recent years. This does not preclude a radical change of behaviour, however any radical change in cessation is outside the current historical experience (National Cancer Institute, 1997).

The good news is recent legal developments and funding of tobacco control programs may provide for increased rates of cessation and thus reductions in smoking prevalence. Continued surveillance is required to monitor changes in smoking to ascertain whether prevalence is in fact stable.

Bibliography

Ades, A. E. and Nokes, D. J. (1993). Modeling age- and time-specific incidence from seroprevalence: toxoplasmosis. *Am J Epidemiol*, 137(9):1022–1034. 79

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag New York, Inc., New York, NY. 47, 51, 58, 60, 61, 122, 161, 250

Anderson, R. M. and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and control.* Oxford University Press, New York, NY. 50, 51, 56

Anderson, R. N. (1999). *United States Life Tables, 1997.* National Vital Statistics Reports, Volume 47. National Center for Health Statistics, Hyattsville, MD. 281

Armitage, P. (1971). Discussion of 'The age distribution of cancer'. *J R Statist Soc A*, 134:155–156. 18, 217

Armitage, P. and Doll, R. (1954). The age distribution of cancer and a multistage theory of carcinogenesis. *Br J Cancer*, 8:1–11. 50, 215

Australian Bureau of Statistics (2000). *Causes of Death, Australia 1999.* Cat. no. 3303.0. Australian Bureau of Statistics, Canberra. 30

Australian Institute of Health and Welfare (1999). *1998 National Drug Strategy Household Survey: First results.* AIHW (Drug Statistics Series), AIHW cat. no. PHE 15. Canberra. 38

Australian Institute of Health and Welfare (2001). *National Health Data Dictionary. Version 10.* Australian Institute of Health and Welfare, AIHW cat. no. HWI 30. Canberra. 42

Australian Institute of Health and Welfare and Australasian Association of Cancer Registries (1999). *Cancer in Australia 1996: Incidence and mortality data for 1996 and selected data for 1997 and 1998.* AIHW cat. no. CAN 7. AIHW (Cancer Series), Canberra. 30

Becker, N. G. and Marschner, I. C. (2001). Advances in medical statistics arising from the AIDS epidemic. *Stat Methods Med Res*, 10(2):117–140. 51

Bennett, S. A. and Magnus, P. (1994). Trends in cardiovascular risk factors in Australia. Results from the National Heart Foundation's Risk Factor Prevalence Study, 1980–1989. *Med J Aust*, 161(9):519–527. 36

Berrino, F., Capocaccia, R., Estève, J., Gatta, G., Hakulinen, T., Micheli, A., Sant, M., and Verdecchia, A. (1999). *Survival of Cancer Patients in Europe: the EUROCARE-2 Study.* International Agency for Research on Cancer, IARC Scientific Publications No. 151, Lyon. 2

Birkett, N. J. (1997). Trends in smoking by birth cohort for births between 1940 and 1975: a reconstructed cohort analysis of the 1990 Ontario Health Survey. *Prev Med*, 26(4):534–541. 49, 120

Bray, I., Brennan, P., and Boffetta, P. (2000). Projections of alcohol- and tobacco-related cancer mortality in Central Europe. *Int J Cancer*, 87(1):122–128. 24

Brenner, H. (1993). A birth cohort analysis of the smoking epidemic in West Germany. *J. Epidemiol. Community Health*, 47(1):54–58. 77

Breslow, N. E. and Day, N. E. (1987). *Statistical Methods in Cancer Research. Volume II: The Design and Analysis of Cohort Studies.* International Agency for Research on Cancer, Lyon. 47, 226

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics (with discussion). *Biometrics*, 42(4):693–734. 56, 227

Brønnum-Hansen, H. (1999). How good is the Prevent model for estimating the health benefits of prevention? *J Epidemiol Community Health*, 53(5):300–305. 251

Brown, C. C. and Chu, K. C. (1987). Use of multistage models to infer stage affected by carcinogenic exposure: example of lung cancer and cigarette smoking. *J Chron Dis*, 40:171S–179S. 18, 216

Brown, K. S. and Forbes, W. F. (1974). A mathematical model of aging processes. *Gerontology*, 29:46–51. 219

Brown, P. N., Byrne, G. D., and Hindmarsh, A. C. (1989). VODE: a variable-coefficient ODE solver. *SIAM J Sci Stat Comput*, 10(5):1038–1051. 85

Burns, D. M., Berowitz, N. L., and Amacher, R. H. (2001). *Risks Associated with Smoking Cigarettes with Low Machine-Measured Yields of Tar and Nicotine. Volume 13.* U. S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 19, 196

Burns, D. M., Garfinkel, L., and Samet, J. M. (1997a). *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control, Monograph 8.* U. S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 194

Burns, D. M., Lee, L., Shen, L. Z., Gilpin, E., Tolley, H. D., Vaughn, J., and Shanks, T. G. (1997b). Cigarette smoking behavior in the United States. In Burns, D. M., Garkinkel, L., and Samet, J. M., editors, *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control, Monograph 8*, book chapter 2, pages 13–112. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 49, 77, 78, 120, 121, 154, 158, 194, 196, 199

Burns, D. M., Shanks, T. G., Choi, W., Thun, M. J., Heath, C. W., and Garkinkel, L. (1997c). The American Cancer Society Cancer Prevention Study I: 12-year followup of 1 million men and women. In Burns, D. M., Garkinkel, L., and Samet, J. M., editors, *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control, Monograph 8*, book chapter 3, pages 113–304. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 96, 100, 111, 226

CDC (1999). Tobacco use — United States, 1900–1999. *MMWR Morb. Mortal. Wkly. Rep.*, 48:986–993. 280

Chiang, C. L. (1968). *Introduction to Stochastic Methods in Biostatistics.* Wiley, New York, NY. 47, 50, 56

Chiang, C. L. (1980). *An Introduction to Stochastic Processes and Their Applications.* Robert E. Kreiger Publishing Co., Huntington, NY. 161

Christie, D., Gordon, I., and Robinson, K. (1986). Smoking in an industrial population. An analysis by birth cohort. *Med J Aust*, 145(1):11–14. 49, 120

Chu, K. C., Baker, S. G., and Tarone, R. E. (1999). A method for identifying abrupt changes in U.S. cancer mortality trends. *Cancer*, 86(1):157–169. 282

Clayton, D. and Schifflers, E. (1987). Models for temporal variation in cancer rates. II: Age-period-cohort models. *Stat Med*, 6(4):469–481. 22

Cleveland, W. S. (1993). *Visualizing Data.* Hobart Press, Summit, NJ. 125

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *J Am Statis Assoc*, 83:596–610. 283

Cochran, W. G. (1977). *Sampling Techniques.* John Wiley and Sons, New York, NY, 3rd edition. 84

Coleman, M. P., Esteve, J., Damiecki, P., Arslan, A., and Renard, H. (1993). *Trends in Cancer Incidence and Mortality*, volume 121 of *IARC Scientific Publications.* International Agency for Research on Cancer, Lyon. 22, 32

Collett, D. (1991). *Modelling Binary Data.* Chapman and Hall, London. 82

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Anal*, 5(4):315–327. 47, 51, 58, 159

Commonwealth Department of Health and Aged Care (1999). *Australia's National Tobacco Campaign. Evaluation Report Volume One.* Commonwealth Department of Health and Aged Care, Canberra. 247

Commonwealth Department of Health and Family Services and Australian Institute of Health and Welfare (1998). *National Health Priority Areas Report on Cancer Control 1997.* Commonwealth Department of Health and Family Services and Australian Institute of Health and Welfare, AIHW Cat. No. PHE 4. Canberra. 20

Cox, B. (1995). *Projections of the Cancer Burden in New Zealand.* Public Health Commission, Wellington. 22

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press, Cambridge. 82, 230, 246

Day, N. E. and Brown, C. C. (1980). Multistage models and primary prevention of cancer. *J Natl Cancer Inst*, 64:977–989. 216

De Angelis, D., Gilks, W. R., and Day, N. E. (1998). Bayesian projection of the acquired immune deficiency syndrome epidemic (with discussion). *Appl Statis*, 47(4):449–498. 50, 250

Department of Statistics (1979). *Cigarette Smoking Bulletin No. 24: 1976 Census of Population and Dwellings*. Department of Statistics, Wellington. 38, 198

Department of Statistics (1983). *Bulletin on Cigarette Smoking: New Zealand Census of Population and Dwellings, 1981*. Department of Statistics, Wellington. 38, 198

Department of Statistics and Department of Health (1992). *Tobacco Statistics 1991*. Department of Statistics and Department of Health, Wellington. 33

Diamond, I. D. and McDonald, J. W. (1991). Analysis of current status data. In Trussel, R., Hankinson, R., and Tilton, J., editors, *Demographic Applications of Event History Analysis*, book chapter 12, pages 231–252. Oxford University Press, Oxford. 79, 121, 172

Doll, R. (1971). The age distribution of cancer: implications for models of carcinogensis (with discussion). *J R Statist Soc A*, 134(133):155. 17, 217

Doll, R. and Hill, A. B. (1966). Mortality of British doctors in relation to smoking: observations on coronary thrombosis. *Natl Cancer Inst Monogr*, 19:205–268. 11, 100

Doll, R., Payne, P., and Waterhouse, J. (1966). *Cancer Incidence in Five Continents: A Technical Report*. UICC, Berlin. 32, 110

Doll, R. and Peto, R. (1978). Cigarette smoking and bronchial carcinoma: dose and time relationships among regular smokers and lifelong non-smokers. *J Epidemiol Community Health*, 32:303–313. 16, 18, 195, 216, 217, 222, 223

Doll, R., Peto, R., Wheatley, K., et al. (1994). Mortality in relation to smoking: 40 years' observations on male British smokers. *BMJ*, 309:901–911. 100, 110

Donnelly, C. A. and Ferguson, N. M. (1999). *Statistical Aspect of BSE and vCJD: Models for epidemics*. Chapman and Hall/CRC, London. 51

Doyle, E. J. (1985). *A Cohort Analysis of Smoking and Lung Cancer in Australia, Canada and the United Kingdom*. Thesis/dissertation, University of Newcastle, NSW, Australia. 24, 211, 213, 221

Dyba, T. and Hakulinen, T. (2000). Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Stat Med*, 19(13):1741–1752. 22

Dyba, T., Hakulinen, T., and Paivarinta, L. (1997). A simple non-linear model in incidence prediction. *Stat Med*, 16(20):2297–2309. 22

Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival Models and Data Analysis.* John Wiley and Sons, New York, NY. 121

English, D. R., Holman, C. D. J., Milne, E., Winter, M., Hulse, G., and Codde, J. (1995). *The Quantification of Drug Caused Morbidity and Mortality in Australia, 1995 edition.* Commonwealth Department of Human Services and Health, Canberra. 2, 95, 100, 110, 288

Enstrom, J. E. (1999). Smoking cessation and mortality trends among two united states populations. *J Clin. Epidemiol*, 52(9):813–825. 250

Estève, J., Benhamou, E., and Raymond, L. (1994). *Statistical Methods in Cancer Research. Volume IV: Descriptive Epidemiology.* International Agency for Research on Cancer, Lyon. 47

Forbes, W. F. and Gibberd, R. W. (1984). Mathematical models of carcinogenesis: A review. *Math Sciences*, 9:95–110. 214, 219

Forey, B. A., Lee, P. N., and Fry, J. S. (1998). Updating UK estimates of age, sex and period specific cumulative constant tar cigarette consumption per adult. *Thorax*, 53(10):875–878. 221

Friedl, H. (1997). On the asymptotic moments of Pearson type statistics based on resampling procedures. *Computation Stat*, 12:265–277. 230

Friedman, G. D., Tekawa, I., Sadler, M., and Sidney, S. (1997). Smoking and mortality: The Kaiser Permanente experience. In Burns, D. M., Garkinkel, L., and Samet, J. M., editors, *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control, Monograph 8*, book chapter 6, pages 477–499. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 100, 113

Gaffney, M. and Altshuler, N. (1988). Examination of the role of cigarette smoking in lung cancer carcinogenesis using multistate models. *J Natl Cancer Inst*, 80:925–931. 50, 217, 219

Gail, M. H., Kessler, L., Midthune, D., and Scoppa, S. (1999). Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality. *Biometrics*, 55(4):1137–1144. 50

Gill, R. D. and Johansen, S. (1990). A survey of product integration with a view toward application in survival analysis. *Ann Statis*, 18(4):1501–1555. 60

Giovino, G. A., Henningfield, J. E., Tomar, S. L., Escobedo, L. G., and Slade, J. (1995). Epidemiology of tobacco use and dependence. *Epidemiol Rev*, 17(1):48–65. 11, 87

Goumas, C., O'Connell, D. L., Smith, D. P., and Armstrong, B. K. (2001). *Lung cancer in New South Wales in 1973 to 1998*. Cancer Council NSW, Sydney. 56

Graham, H. (1996). Smoking prevalence among women in the European community 1950–1990. *Soc Sci. Med*, 43(2):243–254. 14

Green, L. W., Eriksen, M. P., Bailey, L., and Husten, C. (2000). Achieving the implausible in the next decade's tobacco control objectives. *Am J Public Health*, 90(3):337–339. 280

Gunning-Schepers, L. J. (1999). Models: instruments for evidence based policy. *J Epidemiol Community Health*, 53(5):263. 46, 252, 289

Gunning-Schepers, L. J. and Barendregt, J. J. (1992). Timeless epidemiology or history cannot be ignored. *J Clin Epidemiol*, 45:365–372. 21, 251

Hakulinen, T. (1996). The future cancer burden as a study subject. *Acta Oncol.*, 35(6):665–670. 246

Hakulinen, T. and Dyba, T. (1994). Precision of incidence predictions based on Poisson distributed observations. *Stat Med*, 13(15):1513–1523. 230

Hakulinen, T. and Pukkula, E. (1981). Future incidence of lung cancer: Forecasts based on hypothetical changes in smoking habits of males. *Int J Epidemiol*, 10:233–240. 49

Haldorsen, T. and Grimsrud, T. K. (1999). Cohort analysis of cigarette smoking and lung cancer incidence among Norwegian women. *Int J Epidemiol*, 28(6):1032–1036. 25, 213, 221, 223

Harris, J. E. (1980). Patterns of cigarette smoking. In *The Health Consequences of Smoking for Women, a report of Surgeon General*, book chapter 3, pages 15–42. US Government Printing Office, Washington, DC. 120

Harris, J. E. (1983). Cigarette smoking among successive birth cohorts of men and women in the United States during 1900–80. *J Natl Cancer Inst*, 71(3):473–479. 34, 48, 76, 77, 94, 120, 121

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London. 23, 199

Hill, D., White, V., and Letcher, T. (1999). Tobacco use among Australian secondary students in 1996. *Aust N Z J Public Health*, 23(3):252–259. 196, 197

Hill, D., White, V., and Segan, C. (1995). Prevalence of cigarette smoking among Australian secondary school students in 1993. *Aust J Public Health*, 19(5):445–449. 197

Hill, D. J. (1988). Australian patterns of tobacco smoking in 1986. *Med J Aust*, 149(1):6–10. 99

Hill, D. J. and Gray, N. J. (1982). Patterns of tobacco smoking in Australia. *Med J Aust*, 1(1):23–25. 198

Hill, D. J. and White, V. M. (1995). Australian adult smoking prevalence in 1992. *Aust N Z J Pub Health*, 19:305–308. 196

Hill, D. J., White, V. M., Pain, M. D., and Gardner, G. J. (1990). Tobacco and alcohol use among Australian secondary schoolchildren in 1987. *Med J Aust*, 152(3):124–130. 196, 197

Hill, D. J., White, V. M., and Scollo, M. M. (1998). Smoking behaviours of Australian adults in 1995: trends and concerns. *Med J Aust*, 168(5):209–213. 281

Hill, D. J., White, V. M., Williams, R. M., and Gardner, G. J. (1993). Tobacco and alcohol use among Australian secondary school students in 1990. *Med J Aust*, 158(4):228–234. 196, 197

Hill, G. B. (1996). The value of the population attributable risk percentage. *Am J Public Health*, 86(10):1483. 21

Holford, T. R., Zhang, Z., Zheng, T., and McKay, L. A. (1996). A model for the effect of cigarette smoking on lung cancer incidence in Connecticut. *Stat Med*, 15(6):565–580. 25, 213, 221, 223, 226

Holman, C. D. and Armstrong, B. K. (1982). *Cancer Mortality Trends in Australia 1910-1979*. Cancer Council of Western Australia, Perth. 3, 30

Holman, C. D. J., Armstrong, B. K., Arias, L. N., and Martin, C. A. (1990). *The Quantification of Drug Caused Morbidity and Mortality in Australia 1988*. Australian Government Publishing Service, Canberra. 95, 100, 110, 281, 288

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Anal*, 5(3):239–264. 47, 51, 159

Hummer, R. A., Nam, C. B., and Rogers, R. G. (1998). Adult mortality differentials associated with cigarette smoking in the USA. *Pop Res Policy Rev*, 17:285–304. 158

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *J Comput Graphic Statis*, 5(3):299–314. 283

International Agency for Research on Cancer (1986). *Tobacco: A major international health hazard*. IARC Scientific Publications No. 74. International Agency for Research on Cancer, Lyon. 18

Jackson, R. and Beaglehole, R. (1985). Secular trends in underreporting of cigarette consumption. *Am J Epidemiol*, 122(2):341–344. 169, 196

Jee, S. H., Kim, I. S., Suh, I., Shin, D., and Appel, L. J. (1998). Projected mortality from lung cancer in South Korea, 1980–2004. *Int. J Epidemiol*, 27(3):365–369. 22

Jolley, D. and Giles, G. G. (1992). Visualizing age-period-cohort trend surfaces: a synoptic approach. *Int J Epidemiol*, 21(1):178–182. 7

Joly, P., Letenneur, L., Alioum, A., and Commenges, D. (1999). PHMPL: a computer program for hazard estimation using a penalized likelihood method with interval-censored and left-truncated data. *Comput Methods Programs Biomed*, 60(3):225–231. 71, 152

Kahn, D. A. (1966). The Dorn study of smoking and mortality among U.S. veterans: Report on eight and one-half years of observation. *Nat Cancer Inst Monogr*, 19:1–125. 100, 111

Kawachi, I., Colditz, G. A., Stampfer, M. J., Willett, W. C., Manson, J. E., Rosner, B., Hunter, D. J., Hennekens, C. H., and Speizer, F. E. (1997). Smoking cessation and decreased risks of total mortality, stroke, and coronary heart disease incidence among women: A prospective cohort study. In Burns, D. M., Garkinkel, L., and Samet, J. M., editors, *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control, Monograph 8*, book chapter 8, pages 531–565.

U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 100

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Phil Trans R Soc Lond A*, 332:487–509. 6, 47, 61, 74, 121, 125

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion). *J R Statist Soc A*, 154(3):371–412. 47, 56, 78, 79, 121

Keiding, N. (1992). Independent delayed entry. In Klein, J. P. and Goel, P., editors, *Survival Analysis: State of the art*. Kluwer Academic Publishers, Boston. 71

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for censored and truncated data*. Springer-Verlag New York Inc., New York, NY. 71, 74, 124

Knorr-Held, L. and Rainer, E. (2001). Projections of lung cancer mortality in West Germany: A case study in Bayesian prediction. *Biostatistics*, 2:109–129. 24

Korn, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. *Am J Epidemiol*, 145(1):72–80. 71, 122

La Vecchia, C., Levi, F., Decarli, A., Wietlisbach, V., Negri, E., and Gutzwiller, F. (1988). Trends in smoking and lung cancer mortality in Switzerland. *Prev. Med*, 17(6):712–724. 22

LaCroix, A. Z., Lang, J., Scherr, P., Wallace, R. B., Cornoni-Huntley, J., Berkman, L., Curb, J. D., Evans, D., and Hennekens, C. H. (1991). Smoking and mortality among older men and women in three communities. *N Engl J Med*, 324:1619–1625. 100, 113

Langroo, M. K., Wise, K. N., Duggleby, J. C., and Kotler, L. H. (1991). A nationwide survey of 222Rn and gamma radiation levels in Australian homes. *Health Phys.*, 61(6):753–761. 14

Lee, P. N. (1979). *Cigarette smoking and lung cancer. A new mathematical model.* Tobacco Advisory Council, Document TA 1243. London. 18

Lee, P. N. (1995). Studying the relationship of smoking to lung cancer using a multistage model of carcinogenesis: A review. Unpublished Work. 18, 217

Lee, P. N. and Forey, B. A. (1996). Misclassification of smoking habits as a source of bias in the study of environmental tobacco smoke and lung cancer. *Stat. Med*, 15(6):581–605. 55

Lee, P. N. and Forey, B. A. (1998). Trends in cigarette consumption cannot fully explain trends in British lung cancer rates. *J Epidemiol Community Health*, 52(2):82–92. 25, 211, 213, 220, 221

Lindsey, J. K. (2001). *Nonlinear Models in Medical Statistics*. Oxford University Press, Oxford. 23

Loader, C. (1999). *Local Regression and Likelihood*. Springer Verlag New York, Inc., New York, NY. 23, 75, 82, 86, 125, 178, 283

Lubin, J. H., Blot, W. J., Berrino, F., Flamant, R., Gillis, C. R., Kunze, M., Schmahl, D., and Visco, G. (1984). Patterns of lung cancer risk according to type of cigarette smoked. *Int J Cancer*, 33:569–576. 19

Ma, S. and Wong, C. M. (1999). Estimation of prevalence proportion rates. *Int J Epidemiol*, 28(1):175. 82, 83

Malarcher, A. M., Schulman, J., Epstein, L. A., Thun, M. J., Mowery, P., Pierce, B., Escobedo, L., and Giovino, G. A. (2000). Methodological issues in estimating smoking-attributable mortality in the United States. *Am J Epidemiol.*, 152(6):573–584. 96

Mantel, H., Forbes, W. F., Thompson, M. E., and Gibberd, R. W. (1986). Quantitative models of lung cancer mortality. II. Predicting lung cancer mortalities for a population depending on the level of smoking. *Can. J Public Health*, 77(3):208–215. 24, 213, 221

Marang-Van de Mheen, P. J., Smith, G. D., Hart, C. L., and Hole, D. J. (2001). Are women more sensitive to smoking than men? Findings from the Renfrew and Paisley study. *Int J Epidemiol*, 30(4):787–792. 19, 213

Marschner, I. C. (1997). A method for assessing age-time disease incidence using serial prevalence data. *Biometrics*, 53(4):1384–1398. 79

Mathers, C., Vos, T., and Stevenson, C. (1999). *The Burden of Disease and Injury in Australia*. AIHW cat. no. PHE 17. Australian Institute of Health and Welfare, Canberra. 1, 252

Mattson, M. E. and Kessler, L. G. (1987). The use of time-related epidemiologic data on smoking for planning cancer control programs. *J Chron Dis*, 40:25–37. 11

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, New York, 2nd edition. 174, 199, 228, 246

McKeague, I. W. and Utikal, K. J. (1990). Inference for a nonlinear counting process regression model. *Ann Statis*, 18(3):1172–1187. 74

Mendez, D. and Warner, K. E. (2000). Smoking prevalence in 2010: why the Healthy People goal is unattainable. *Am J Public Health*, 90(3):401–403. 46, 280, 290

Mendez, D., Warner, K. E., and Courant, P. N. (1998). Has smoking cessation ceased? Expected trends in the prevalence of smoking in the United States. *Am J Epidemiol*, 148(3):249–258. xix, 46, 49, 54, 94, 151, 158, 172, 173, 179, 280, 281, 282, 285, 286, 287, 288

Miller, A. B. (1999). Tobacco and cancer: what has been, and could be, achieved? *Cancer Strategy*, 1:165–169. 5

Ministry of Health (1999a). *Progress on Health Outcome Targets 1999.* Ministry of Health, Wellington. 281

Ministry of Health (1999b). *Taking the Pulse: The 1996/97 New Zealand Health Survey.* Ministry of Health, Wellington. 41, 57

Ministry of Health (2000). *Progress on Health Outcome Targets 2000.* Ministry of Health, Wellington. 6, 41

Ministry of Health (2001). *Tobacco Facts 2001.* Ministry of Health, Wellington. 33

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill, New York, NY, 3rd edition. 68, 75

Moolgavkar, S. H., Dewanji, A., and Luebeck, G. (1989). Cigarette smoking and lung cancer: Reanalysis of the British doctors' data. *JNCI*, 81:415–420. 19, 50, 217, 218, 219, 220

Moolgavkar, S. H., Lee, J. A. H., and Stevens, R. G. (1998). Analysis of vital statistics data. In Rothman, K. J. and Greenland, S., editors, *Modern Epidemiology*, book chapter 24, pages 481–497. Lippincott-Raven, Philadelphia, PA, 2nd edition. 22, 24, 122

Moolgavkar, S. H. and Luebeck, E. G. (1992). Multistage carcinogenesis: population-based model for colon cancer. *J Natl Cancer Inst*, 84(8):610–618. 247

Moolgavkar, S. H. and Luebeck, G. (1990). Two-event model for carcinogenesis: biological, mathematical, and statistical considerations. *Risk Anal*, 10(2):323–341. 247

Mooney, G., Irwig, L., and Leeder, S. (1997). Priority setting in health care: unburdening from the burden of disease. *Aust N Z J Public Health*, 21(7):680–681. 21

Morgenstern, H. and Bursic, E. S. (1982). A methods for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *J Commun Health*, 7:292–309. 21, 251

Moulton, L. H. and Zeger, S. L. (1991). Bootstrapping the generalized linear-model. *Comput Stat Data Analysis*, 11:53–63. 230

Müller, H. G. and Wang, J. L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50(1):61–76. 124

Mullins, R., Hill, D., and Borland, R. (2000). Changing the way smoking is measured among Australian adults. In Trotter, L. and Letcher, T., editors, *Quit Evaluation Studies No. 10, 1998–1999*, book chapter 5, pages 59–66. Victorian Smoking and Health Program, Melbourne. 36, 42

Murray, C. J. and Lopez, A. D. (1997a). Global mortality, disability, and the contribution of risk factors: Global burden of disease study. *Lancet*, 349(9063):1436–1442. 21

Murray, C. J., Lopez, A. D., and Jamison, D. T. (1994). The global burden of disease in 1990: summary results, sensitivity analysis and future directions. *Bull. World Health Organ*, 72(3):495–509. 21

Murray, C. J. L. and Lopez, A. D. (1997b). Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *The Lancet*, 349:1436–1442. 1, 51

Nam, C. B., Roger, R. G., and Hummer, R. A. (1996). Impact of future cigarette smoking scenarios on mortality of the adult population in the United States, 2000–2050. *Soc Biol*, 43:155–168. 22

National Cancer Institute (1997). *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control, Volume 8.* U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 94, 288, 289, 290

National Center for Health Statistics (2000). *1997 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description.* National Center for Health Statistics, Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, Hyattsville, MD. 281

New Zealand Health Information Service (2000a). *Cancer: New Registrations and Deaths 1996.* Ministry of Health, Wellington. 30

New Zealand Health Information Service (2000b). *Mortality and Demographic Data 1997.* Ministry of Health, Wellington. 30

Ogata, Y., Katsura, K., Keiding, N., Holst, C., and Green, A. (2000). Empirical Bayes age-period-cohort analysis of retrospective incidence data. *Scand J Statis*, 27:415–432. 153, 250

Patrick, D. L., Cheadle, A., Thompson, D. C., Diehr, P., Koepsell, T., and Kinne, S. (1994). The validity of self-reported smoking: a review and meta-analysis. *Am J Public Health*, 84(7):1086–1093. 11

Pearce, N. (1996). Traditional epidemiology, modern epidemiology, and public health. *Am J Public Health*, 86(5):678–683. 253

Peto, R. (1977). Epidemiology, multistage models, and short-term mutagenicity tests. In Hyatt, H., Watson, J., and Winsten, J. A., editors, *Origins of Human Cancer*, pages 1403–1428. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. 14, 20

Peto, R. (1986). Influence of dose and duration of smoking on lung cancer rates. In Zaridze, D. G. and Peto, R., editors, *Tobacco: A major international health hazard*, pages 23–33. International Agency for Research on Cancer, Lyon. 221

Peto, R., Lopez, A. D., Boreham, J., et al. (1994). *Mortality from Smoking in Developed Countries 1950–2000.* Oxford University Press, Oxford. 2, 94

Peto, R., Lopez, A. D., Boreham, J., Thun, M., and Heath, C. J. (1992). Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet*, 339(8804):1268–1278. 96

Pierce, J. P., Aldrich, R. N., and S, H. (1987). Uptake and quitting smoking trends in Australia 1974-1984. *Prev Med*, 16. 86

Pike, M. C. (1966). A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, 22:142–161. 215

Prescott, E., Osler, M., Andersen, P. K., Hein, H. O., Borch-Johnsen, K., Lange, P., Schnohr, P., and Vestbo, J. (1998). Mortality in women and men in relation to smoking. *Int J Epidemiol*, 27(1):27–32. 19, 100, 112, 213

Public Health Commission (1994). *Tobacco Products: The Public Health Commission's advice to the Minister of Health 1993–1994*. Public Health Commission, Wellington. 20

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann Statis*, 11(2):453–466. 73

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 2nd edition. 228, 283, 284

Rao, J. N. and Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics*, 48(2):577–585. 81, 84

Risk Factor Prevalence Study Management Committee (1990). *Risk Factor Prevalence Study: Survey No. 3 1989*. National Heart Foundation of Australia and Australian Institute of Health, Canberra. 36

Robertson, C. and Boyle, P. (1998). Age-period-cohort models of chronic disease rates. II: Graphical approaches. *Stat Med*, 17(12):1325–1339. 7

Rogers, R. G. and Powell-Griner, E. (1991). Life expectancies of cigarette smokers and nonsmokers in the United States. *Soc Sci. Med.*, 32(10):1151–1159. 94, 158, 281

Rosenbaum, W. L., Sterling, T. D., and Weinkam, J. J. (1998). Use of multiple surveys to estimate mortality among never, current, and former smokers: changes over a 20-year interval. *Am J Public Health*, 88(11):1664–1668. 100

Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology*. Lippincott-Raven, Philadelphia, 2nd edition. 81, 95, 107

Samet, J. M. (1994). *Epidemiology of Lung Cancer.* Lung Biology in Health and Disease. Marcel Dekker, Inc., New York, NY. 14, 19, 211, 216

Saracci, R. (1987). The interactions of tobacco smoking and other agents in cancer etiology. *Epidemiol Rev*, 9:175–193. 14

SAS Institute Inc. (1999). *SAS/STAT User's Guide, Version 8.* SAS Institute Inc., Cary, NC. 160, 282

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression.* John Wiley and Sons, New York, NY. 169

Shah, B. V., Barnwell, B. G., and Bieler, G. S. (1996). *SUDAAN User's Manual, Release 7.0.* Research Triangle Institute, Research Triangle Park, NC. 282

Shyrock, H. S., Siegel, J. S., and Stockwell, E. G. (1976). *The Methods and Materials of Demography.* Studies in Population. Academic Press, Orlando, Florida. 32, 47, 56, 169, 180

Skov, T., Deddens, J., Petersen, M. R., and Endahl, L. (1998). Prevalence proportion ratios: estimation and hypothesis testing. *Int J Epidemiol*, 27(1):91–95. 82, 83, 282

Smith, G. and Ebrahim, S. (2001). Epidemiology — is it time to call it a day? *Int J Epidemiol*, 30(1):1–11. 2

Sorlie, P. D., Kannel, W. D., and O'Connor, G. (1989). Mortality associated with respiratory function and symptoms in advanced age. The Framingham Study. *Am Rev Respir Dis*, 140:S49–S55. 100, 112

Stevens, R. G. and Moolgavkar, S. H. (1984). A cohort analysis of lung cancer and smoking in British males. *Am J Epidemiol*, 119(4):624–641. 18, 24, 211, 213, 221, 222

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, Princeton, NJ. 85

Supramaniam, R., Smith, D., Coates, M., and Armstrong, B. (1999). *Survival from cancer in New South Wales in 1980 to 1995.* NSW Cancer Council, Sydney. 2

Swartz, J. B. (1992). Use of a multistage model to predict time trends in smoking induced lung cancer. *J Epidemiol Community Health*, 46(3):311–315. 25, 211, 213, 220, 221

Taylor, R. and McNeil, D. (1997). *Projections of Incidence of Major Cancers in NSW to 2001*. NSW Cancer Council, Sydney. 22

Taylor, R. J., Morrell, S. L., Mamoon, H. A., and Wain, G. V. (2001). Effects of screening on cervical cancer incidence and mortality in new south wales implied by influences of period of diagnosis and birth cohort. *J Epidemiol Community Health*, 55(11):782–788. 22

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox model*. Springer-Verlag, New York, NY. 75, 124, 125

Thomas, D. C. (1987). Pitfalls in the analysis of exposure-time-response relationships. *J Chronic. Dis*, 40 Suppl 2:71S–78S. 18

Thomas, D. C. (1988). Models for exposure-time-response relationships with applications to cancer epidemiology. *Annu. Rev Public Health*, 9:451–482. 214, 220, 221, 246

Thun, M. J., Day-Lally, C., Myers, D. G., Calle, E. E., Flanders, W. D., Zhu, B., Namboodiri, M. M., Heart, and Heath, C. W. (1997). Trends in tobacco smoking and mortality from cigarette use in Cancer Prevention Studies I (1959 through 1965) and II (1982 through 1988). In Burns, D. M., Garkinkel, L., and Samet, J. M., editors, *Changes in Cigarette-related Disease Risks and Their Implication for Prevention and Control, Monograph 8*, book chapter 4, pages 305–382. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. xvi, 9, 14, 16, 17, 94

Thun, M. J. and Heath, C. W. J. (1997). Changes in mortality from smoking in two American Cancer Society prospective studies since 1959. *Prev Med*, 26(4):422–426. 100, 111, 226

Todd, G. F. (1975). *Changes in Smoking Patterns in the UK. Ocassional paper 1.* Tobacco Research Council, London. 211

Todd, G. F. (1978). Cigarette consumption per adult of each sex in various countries. *J. Epidemiol. Community Health*, 32(4):289–293. 9, 94

Tolley, H. D., Crane, L., and Shipley, N. (1991). Smoking prevalence and lung cancer deaths rates. In Shopland, D. R., Burns, D. M., Samet, J. M., and Gritz, E. R., editors, *Strategies to Control Tobacco Use in the United States: a blueprint for* *public health action in the 1990's, Monograph 1*, book chapter 3, pages 73–144. U.S. Department of Health and Human Services, Public Health Service, National Institutes of Health, National Cancer Institute, Bethesda, MD. 24, 25, 49, 56, 57, 213, 220, 252

Tolley, H. D. and Manton, K. G. (1991). Intervention effects among a collection of risks. *Trans Soc Actuaries*, 43:443–468. 48, 49, 56

Townsend, J. L. (1978). Smoking and lung cancer: a cohort data study of men and women in England and Wales 1935–70. *J R Statist Soc A*, 141:95–107. 24, 211, 213, 220, 226

Tyrrell, I. R. (1999). *Deadly enemies: tobacco and its opponents in Australia.* UNSW Press, Sydney. 7, 203, 210, 211

U.S. Census Bureau (2000a). *(NP-D1-A) Projections of the Resident Population by Age, Sex, Race, and Hispanic Origin: 1999 to 2100.* U.S. Census Bureau, Washington, DC. 281

U.S. Census Bureau (2000b). *Resident Population Estimates of the United States by Age and Sex: April 1, 1990 to July 1, 1999, with Short-Term Projection to April 1, 2000.* U.S. Census Bureau, Washington, DC. 281

U.S. Department of Health and Human Services (1989). *Reducing the Health Consequences of Smoking: 25 Years of Progress. A Report of the Surgeon General.* DHHS Publication No. (CDC) 89-8411. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Washington, DC. 86, 93

U.S. Department of Health and Human Services (1990). *The Health Benefits of Smoking Cessation. A Report of the Surgeon General.* DHHS Publication No. (CDC) 90-8416. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control, Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Washington, DC. 19, 87

U.S. Department of Health Education and Welfare (1979). *Smoking and Health. A Report of the Surgeon General.* DHEW Publication No. (PHS) 79-50066. U.S. Department of Health, Education, and Welfare, Public Health Service, Office of the Assistant Secretary for Health, Office on Smoking and Health, Washington, DC. 2

van de Mheen, P. J. and Gunning-Schepers, L. J. (1994). Reported prevalences of former smokers in survey data: the importance of differential mortality and misclassification. *Am J Epidemiol*, 140(1):52–57. 54, 106, 121, 158

van de Mheen, P. J. and Gunning-Schepers, L. J. (1996). Differences between studies in reported relative risks associated with smoking: an overview. *Public Health Reports*, 111(5):420–426. 94, 106

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*. Statistics and Computing. Springer-Verlag New York, Inc., New York, NY. 96, 162

Verdecchia, A., Mariotto, A., Capocaccia, R., Gatta, G., Micheli, A., Sant, M., and Berrino, F. (2001). Incidence and prevalence of all cancerous diseases in Italy: trends and implications. *Eur J Cancer*, 37(9):1149–1157. 50

Walker, A. (2000). Distributional impact of higher patient contributions to australia's pharmaceutical benefits scheme. *Aust Health Rev*, 23(2):32–46. 48

Wall, S. (1995). Epidemiology for prevention. *Int. J Epidemiol*, 24:655–664. 253

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London. 74, 75

Warner, K. E. (1978). Possible increases in underreporting of cigarette consumption. *J Am Statis Assoc*, 73:314–318. 11, 196

Whittemore, A. S. (1988). Effect of cigarette smoking in epidemiological studies of lung cancer. *Stat Med*, 7(1-2):223–238. 18, 223, 226

Whittemore, A. S. and Keller, J. B. (1978). Quantitative theories of carcinogenesis. *SIAM Review*, 20(1):30. 215

Wilcox, H. B., Schoenberg, J. B., Mason, T. J., Bill, J. S., and Stemhagen, A. (1988). Smoking and lung cancer: risk as a function of cigarette tar content. *Prev Med*, 17:263–272. 19

Wilkenfeld, J., Henningfield, J., Slade, J., Burns, D., and Pinney, J. (2000). It's time for a change: cigarette smokers deserve meaningful information about their cigarettes. *J Natl Cancer Inst*, 92(2):90–92. 19, 57, 196

Winstanley, M., Woodward, S., and Walker, N. (1995). *Tobacco in Australia, Facts and Issues, 1995*. Quit Victoria, Australia, 2nd edition. 9, 33, 211

Wolfson, M. C. (1994). POHEM – a framework for understanding and modelling the health of human populations. *Wld Hlth Statis Quart*, 47:157–176. 48

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *J R Statist Soc B*, 62:413–428. 228

Woodward, S. D. (1984). Trends in cigarette consumption in Australia. *Aust N Z J Med*, 14:405–407. 203

Xu, Z., Armstrong, B. K., Blundson, B. J., Rogers, J. M., Musk, A. W., and Shilkin, K. B. (1985). Trends in mortality from malignant mesothelioma of the pleura, and production and use of asbestos in Australia. *Med. J. Aust.*, 143(5):185–187. 14, 31

Yamaguchi, N., Mochizuki-Kobayashi, Y., and Utsunomiya, O. (2000). Quantitative relationship between cumulative cigarette consumption and lung cancer mortality in Japan. *Int J Epidemiol*, 29:963–968. 18, 25