Predicting the future burden of cancer on society.

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester.

by

Mark John Rutherford BSc MSc

Department of Health Sciences

University of Leicester.

Submitted September 14th, 2011.

Predicting the future burden of cancer on society.

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester.

by

Mark John Rutherford BSc MSc

Department of Health Sciences

University of Leicester.

Submitted September 14th, 2011.

Abstract

Predicting the future burden of cancer on society.

M. J. Rutherford.

Evaluating the burden of cancer on society is of great interest to health officials and planning authorities. It is of particular importance to be able to correctly estimate the burden of cancer in the coming years in order that appropriate provisions can be put in place. The vast majority of developed, and also developing, countries have a cancer registry set-up and have at least 20 years of complete data. In the leading developed countries, the cancer registry data is complete and reliable for the past 50 years. Using this data it is possible to estimate key quantities that can be used to assess the burden of cancer.

Prevalence gives a good proxy for the burden of cancer on society; it gives an estimate of the number of people who are alive having had a previous cancer diagnosis. Prevalence can be estimated by combining models for incidence and patient survival. To accurately model the prevalence, it is important to develop the best methods for modelling the incidence and patient survival from population-based cancer registries. Therefore, as part of this thesis, novel methods have been developed for projecting cancer incidence into the future using an approach that treats the data continuously. Also, methods for projecting cancer patient survival have been assessed and improved as part of the work by effectively estimating the quantities in continuous time. These projected estimates have been combined to give future estimates of cancer prevalence.

Making predictions is obviously fraught with danger and, therefore, it should be made clear that these projections are liable to be uncertain and based on strong assumptions. However, if the assumptions of these models are fully understood, they may well provide a useful tool for health and financial planning in terms of assessing the disease burden due to the differing forms of cancer.

Acknowledgements

Firstly, I would like to thank all of the people that have helped or supported me in any way during the process of writing this thesis. I would, however, like to make a special thanks to Dr. Paul Lambert who, in acting as my main supervisor, has invested his time and expertise to ensure that I performed to the very best of my ability. Thanks also goes to the other members of academic staff in the Department of Health Sciences for their help and guidance; particularly to Professor John Thompson for his role as my second supervisor. Further thanks go to the past and current occupants of Room 211 (including honorary members) who have supported me and made the entire process bearable! Michael Crowther deserves extra thanks for reading drafts of the key chapters and providing some insightful comments. Natalie Rutherford has also read drafts of the thesis and sorted out some of my grammatical errors and "typos". I would also like to thank the remainder of my family and friends for their care and support.

In his role as scientific advisor for the PhD thesis, Dr. Paul Dickman has been involved in one of my publications and has provided both support and guidance. He also gave me an opportunity to be involved in teaching on a course which helped to disseminate my `apcfit` software. I would like to thank Enzo Coviello for some interesting e-mail discussions over the 3 years that have helped me to concentrate on some key issues. I would also like to thank Enzo for testing my Stata code, and being the first to publish applied work having used `apcfit`. Dr. Freddie Bray has also been a keen user of `apcfit`, and has helped to improve the software with some of his suggestions. Dr. Bendix Carstensen deserves thanks for being involved in helping to improve the Stata Journal article draft despite his loyalty to R.

Finally, I would like to express my gratitude to the Finnish Cancer Registry for allowing me to use their data for the analyses conducted as part of this thesis; particular thanks go to the director, Professor Timo Hakulinen, for ensuring I received the data promptly and for his input when reading over the work published using the data.

Contents

List of Tables

List of Figures

List of Abbreviations

AC: Age-Cohort.

AIC: Akaike Information Criterion.

AP: Age-Period.

APC: Age-Period-Cohort.

BIC: Bayesian Information Criterion.

df: Degrees of Freedom.

GLM: Generalised Linear Model.

IARC: International Agency for Research on Cancer.

LRT: Likelihood Ratio Test.

NHS: National Health Service.

pdf: Probability Density Function.

PY: Person-Years.

RS: Relative Survival.

SEER: Surveillance Epidemiology and End Results.

UK: United Kingdom.

US: United States (of America).

WHO: World Health Organisation.

CHAPTER 1

Introduction

1.1. Aims of the Thesis

The primary aim of the thesis is to improve methods for estimating the burden of cancer on society and to provide appropriate projections of this estimate. Cancer burden has been used as an all-encompassing term, and also has been defined more specifically as, for example, the prevalence or incidence of the disease. Cancer burden could comprise of a number of different constituents; incidence, survival, mortality, prevalence, cost of care, and quality of life. Through improving the estimation of the components that comprise the burden of cancer on society, the estimation of the burden of cancer on society can be improved. In this thesis, prevalence will be used as a proxy for cancer burden. Models for incidence and survival will be developed; these measures will be appropriately combined to give an estimate of prevalence. Methods to project both incidence and survival will also be developed in order to then provide a projected estimate of prevalence.

1.2. Cancer Burden

Cancer is a term used for a group of diseases that are caused by the uncontrolled growth of abnormal cells. There are over 100 different types of cancer. It is estimated that 7.6 million people in the world died of cancer in 2007 [WHO]. In the UK, cancer is responsible for 126,000 deaths every year [NHS]. It is clear that cancer is a serious health problem across the world, and that the burden that cancer has on society is a heavy one.

The cancer burden is comprised of a number of different components; such as incidence, survival, financial cost, social cost and prevalence. In order to give a true measure of the burden of cancer, it is essential that as many of these elements as possible are taken into account. A number of these quantities can be separately modelled from population-based cancer registry data; such as incidence and survival. However, it should theoretically be possible to give an overall estimate of burden by appropriately combining these quantities. In this thesis,

prevalence will be used as a proxy for the cancer burden. The prevalence of cancer will be estimated from the combination of incidence and survival to estimate the number of people in society with a diagnosis of cancer. This could be used as a starting point for including details on cost and quality of life should this information be available. Cost will be given attention in the discussion but not in any of the chapters including analyses. Health and planning authorities require an estimate of the future number of patients that will need to receive treatment following a diagnosis of cancer. Both the future number of new cancer cases (incidence) and the future number of patients who have had a cancer diagnosis (prevalence) will be of interest when considering the future planning of services. Therefore, it is these measures that will be considered when attempting to provide a future burden estimate for cancer.

1.3. Cancer Registries

The data that are used to estimate both incidence and survival from cancer are generally collected by regional/national cancer registries. In most countries, this data is then compiled to give nationwide data on all registrations of a case of cancer in the population. This data, coupled with population statistics for the country, is sufficient to model incidence and mortality rates of the disease. Cancer registries play an essential role in the improvement of health and healthcare through the monitoring of information on cancer [Thames Cancer Registry]. However, there are only a limited number of variables that can be collected within this infrastructure, and time-varying covariates, such as current treatment, are typically not available in this setting. Information on age, gender and stage are often recorded at diagnosis. The cancer registry then links to the death registry data in order to record if and when the cancer patient has died. Often, the cause of death information is also collected for the cancer patients, and an attempt to ascertain whether or not their death was due to cancer is often made. Cause of death information has been shown to be unreliable in a number of settings [Flanders, 1992; James and Bull, 1996; Satariano et al., 1998], particularly for the oldest age-groups [Hoel et al., 1993; Modelmog et al., 1992].

1.4. Incidence

In order to build towards a complete analysis for predicting the cancer burden into the future, it is necessary to find the most appropriate models for modelling incidence and survival.

Cancer incidence measures the number of new cancer cases in a given time period, and is given as a rate with the denominator comprising of the total population time at risk over that time period. Cancer incidence often varies by key covariates such as age and sex. Understanding changes in the rate of incidence over time and age is vital for assessing disease burden. Models for cancer incidence will be covered in Chapter 2. Further details on the models, and also methods for projecting cancer incidence will be given in Chapters 3, 4 and 5.

1.5. Survival

Survival analysis will be formally introduced in Chapter 6. Data on patient survival are an invaluable tool in the evaluation of progress against cancer [Dickman and Adami, 2006]. The simplest measure to give in a survival analysis setting is overall, or all-cause, survival. That is, all deaths are treated as an event and the proportion of the cohort of patients still alive at a given point in time is evaluated; with censoring appropriately accounted for. However, between different populations, and within populations (due to the effect of age), there are differing risks in the background mortality rate; that is deaths from causes other than cancer. Therefore, a measure that is independent of the background probability of deaths from other causes is sought by cancer registries.

Net survival is a purely hypothetical measure that is independent of the background risk of death. It gives the probability of dying from the disease of interest in the absence of death from other causes. There are two established methods for estimating this quantity; both of which require assumptions under which they estimate net survival. The two measures are cause-specific survival and relative survival. Comparisons of the assumptions required for both measures for use in population-based studies have recently been evaluated [Sarfati et al., 2010]. Both approaches require an independence between the mortality due to cancer and the mortality due to other causes. Cause-specific survival requires cause-of-death information to be accurately recorded. Particularly in the oldest age-groups, it has been shown that cause-of-death information can be inaccurately recorded [Hoel et al., 1993; Modelmog et al., 1992]. Therefore, relative survival methods have become the standard approach to estimating net survival for population-based studies.

Relative survival is the most commonly reported measure of cancer patient survival by cancer registries. Relative survival methods are used to try to obtain an estimate of net survival;

that is, the probability of surviving the disease of interest in the absence of death from other causes. This is a hypothetical measure, but is useful for comparisons between groups in that it is adjusted for the fact that different populations may have different levels of background risk of death. The expected (or background) mortality is normally obtained from nationwide or regional population mortality figures, which are obtained in yearly intervals for age and calendar time, whilst also being estimated separately for each sex. Relative survival as a function of time, $R(t)$, is defined as:

$$R(t) = \frac{S(t)}{S^*(t)}, \tag{1.1}$$

where $S^*(t)$ is the background survival in the population, and $S(t)$ is the observed survival for the cancer patients. A comparison of approaches for estimating relative survival will be given in Chapter 7. The calculation of relative survival is vital for the concept of cancer burden as outlined above. When patients are removed from the cohort of cancer patients through death, the estimate of prevalence must be appropriately adjusted.

A related concept to that of relative survival is that of statistical cure. A cohort of patients is considered to be statistically cured when the excess hazard associated with the disease of interest decreases to zero and the hazard of mortality in the cohort of patients is thus the same as the general population. This corresponds to the fraction given in Equation (1.1) reaching a plateau. Again, this is an important concept when considering the burden of cancer on society and statistical cure will be fully introduced in Section 6.8. When a cancer patient is no longer at an excess risk from their diagnosis of cancer, we can consider removing them from the prevalent cancer cases if interest lies in the burden of cancer.

1.6. Prevalence

Methods exist to accurately model incidence, mortality and survival using population-based cancer registry data. These estimates can then be combined to give an estimate of the prevalence of cancer at a given point in time. The prevalence of any given cancer gives a measure of the number of people in the population that are currently defined as having an active case of cancer. Therefore, prevalence can be considered as a good proxy for the overall cancer burden. Provided that the information is available, it is possible to assess the prevalence

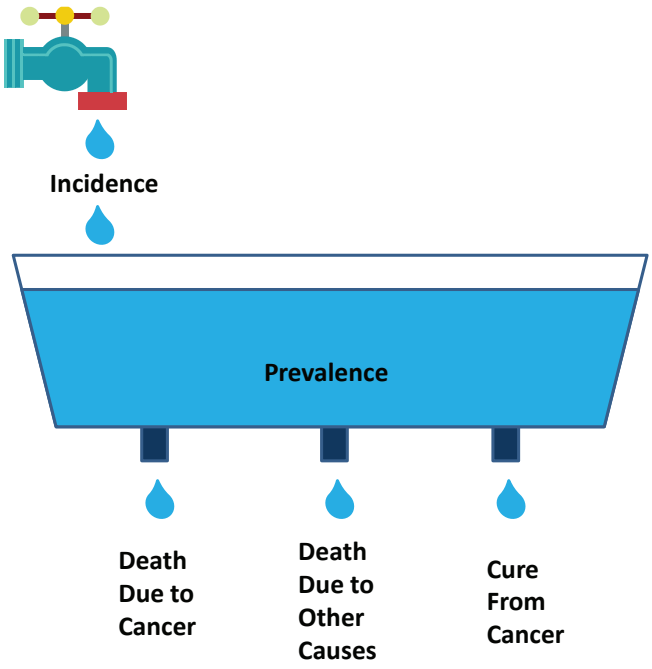

FIGURE 1.1. Simple illustration of the inter-relations for the components of the cancer burden

levels within certain subgroups of the population by adding covariates into the relevant models. Theoretically it should be possible to model the prevalence using a joint model for incidence and survival rather than combining the results of two separate models. This would be of benefit in trying to assign appropriate confidence intervals to the prevalence estimates, which are clearly subject to variability and uncertainty. Estimates of prevalence are considered in Chapter 9, after models for survival and incidence have been developed over the preceding chapters.

Figure 1.1 gives an illustration of the inter-relations between the various measures associated with the reporting of cancer burden. Considering the prevalence to be the level of the water in the bath, new instances of cancer diagnosis can be considered as the water being added to the bath from the tap. The three "plugholes" are then dealt with by the various measures of survival. The first two "plugholes" could be combined to consider just the death of those diagnosed with cancer from any given cause. However, interest often lies in distinguishing

whether or not cancer has been the cause of death. Cause-specific and relative survival analyses are the methods that are mainly used in order to do this. This requires applying competing risks methodology [Putter et al., 2007] to estimate the crude probabilities of death [Lambert et al., 2010]. The final "plughole" relates to those patients that are no longer at risk of the given cancer and can, therefore, for the sake of cancer burden be considered cured. Due to the lack of longitudinal data available to the cancer registries, it is often very difficult to establish the time-to-cure for any given patient. Measures that take a population viewpoint for cure have been developed [De Angelis et al., 1999; Lambert et al., 2007]. However, there are also limitations in these methods for estimating the actual time-to-cure as they employ techniques that assume the asymptote for the cure proportion is reached at an infinite timepoint in the future.

The illustration of the bathwater is a useful one as it helps to consider what the effect of a specific intervention would be for a given cancer. For example, if a new intervention were to have the effect of reducing the deaths due to cancer, this would have an effect not only for the level of the bathwater, but may also consequently increase the flow from the other two "plugholes". That is, an intervention that extends a given patient's life from the cancer death could lead to them dying from something other than their cancer first.

1.7. Projection

Health and planning officials need to plan future treatment and care for the population. It is therefore of interest and importance that accurate projections of the future cancer burden can be obtained from the currently available data. Consideration must be given to methods of projecting incidence and survival should an estimate of future burden be required. The simplest assumption would be to assume that the given values remain the same from the last observation point. However, on the basis that trends have been modelled temporally for the quantities of interest, it is often evident that this assumption would be somewhat naïve in most cases. The other consideration is that projecting an overly complex function into the future would also be fraught with danger. Considering that there is no available information for the future on which to base the shape of the function, it is often considered that in most cases a simple, often linear (at least on a given scale), projection is the most suitable assumption over the short-term [Møller et al., 2003; Bray and Møller, 2006].

Projections for incidence models are considered in Chapter 4, whereas the corresponding chapter for projecting survival is Chapter 8. The two projection methods are combined when used for projecting estimates of prevalence in Chapter 9.

1.8. Layout of Thesis

In Chapter 2, age-period-cohort models are used for modelling incidence. The theory is introduced and an example of the output is given. An approach using restricted cubic splines is fully explored; which gives a more realistic and smooth illustration of the changes of incidence over age, and calendar time. Methods for fitting interactions with other key covariates are proposed and an improvement is suggested using a "reduced" set of spline variables. The output is produced by a user-written package [Rutherford et al., 2010] that has been developed for the statistical software Stata [StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP]. The next chapter looks at the issue of knot placement when using restricted cubic splines in the age-period-cohort setting. A criticism of using splines to model continuous covariates is the choice and placement of the knots. A simulation is undertaken to investigate the optimal methods for selecting the number and placement of the knots. The aim of the chapter is to highlight the insensitivity of the model fit to the choice of knots under the proviso that a sensible approach is adopted. In Chapter 4, a new method for projecting cancer incidence using age-period-cohort models is compared and contrasted to established approaches. A retrospective analysis is used for Finnish cancer data to compare the projections made to the known data. This is possible due to the length and quality of the Finnish registry data. The incidence projection method proposed in Chapter 4 is the topic of a paper currently under peer-review [Rutherford et al., 2011b]. In the fifth Chapter, an investigation is made into the dangers of making projections. Methods to lessen the danger are discussed.

In Chapter 6, a move away from incidence to survival is undertaken. However, the data used are very similar to that used in the earlier chapter and the analyses can all be carried out on the standard data collected by cancer registries. Survival is a key measure involved in estimating the burden of cancer on society, and Chapter 6 introduces the main concepts of registry data analysis; including relative survival, period analysis, and statistical cure. The flexible parametric survival model is defined, and examples are used to show how each of the measures listed can be carried out in the same framework. The seventh chapter describes

a simulation that was carried out in order to investigate the various methods for estimating relative survival in a range of scenarios. The findings of the simulation further emphasise the necessity to making some adjustment for age when calculating the measures associated with cancer burden, even when an overall measure for all ages combined is required. The work in this chapter has recently been published [Rutherford et al., 2011a]. In Chapter 8, methods for projecting survival into the future are investigated in a retrospective analysis in a similar way to that carried out for incidence projections in Chapter 4. The methods that are compared are based largely on the idea of modelled period analysis.

In Chapter 9, the combination of the estimates from the survival projections and the incidence projections are discussed. The estimates from the earlier chapters are combined in order to give an estimate of prevalence. The methods are again examined using a retrospective analysis. A discussion is given in this chapter around the idea of putting a level of uncertainty on the estimates that are projected. The thesis is concluded in Chapter 10 with a general discussion of the work, and a discussion of potential future work in the area.

CHAPTER 2

Age-Period-Cohort models

2.1. Chapter Outline

In this chapter, an introduction to modelling cancer incidence is given. The concept of age-period-cohort models is introduced and a method using restricted cubic splines is fully examined and further developed through an example using Finnish cancer registry data. Software developed as part of the thesis [Rutherford et al., 2010] is used to conduct the analyses and to produce the graphical output.

2.2. Introduction

To construct an estimate of the overall burden of cancer, a measure of the number of new cases needs to be obtained. An Age-Period-Cohort (APC) model provides a modelling tool that can be used to model the incidence/mortality rate of the disease using data routinely collected by cancer, and other disease, registries. The models provide a framework for estimating the trends of disease over the range of the three variables of interest; age, period and cohort. However, APC models suffer from an identifiability problem. The date of birth can be calculated directly from the age at diagnosis (or death) and the date of diagnosis (or death). If fitted directly in a Generalised Linear Model (GLM), this leads to over-parameterisation and, consequently, the exclusion of one of the terms. It is, therefore, necessary to fit constraints to the model in order to extract an identifiable solution for the parameters.

2.3. General Form of the APC Model

The general form of the age-period-cohort model (with $\mathbf{a} = \mathbf{p} - \mathbf{c}$) can be given as:

$$\ln\{\lambda(\mathbf{a}, \mathbf{p})\} = f(\mathbf{a}) + g(\mathbf{p}) + h(\mathbf{c}), \tag{2.1}$$

where f, g and h are functions, and $\mathbf{a}$, $\mathbf{p}$ and $\mathbf{c}$ are the values of age, period and cohort respectively. This model can be used to predict the incidence or mortality rate for any given

combination of age and period. However, due to the direct relationship between the terms, $\mathbf{c} = \mathbf{p} - \mathbf{a}$, the components of this model cannot be uniquely determined. The model needs to be constrained in some way in order to ensure that three functions showing the age, period and cohort effects can be extracted. It has become common to select a step function for the functions f, g and h, and fit a factor model to the data, with the age and calendar timescales split into 5 year intervals [Zheng et al., 1992; Bergström et al., 1996; Cayuela et al., 2004; Gordon et al., 2011].

2.4. Lexis Diagrams

A Lexis diagram allows the full visualisation of the data being analysed [Keiding, 1990]. Figure 2.1 gives an example for a small number of patients. The diagonal lines start at the period of diagnosis on the x-axis, and the age at diagnosis on the y-axis. The line then continues as calendar time progresses and the patient ages. This line is indicative of the follow-up time that is analysed during survival studies (survival analysis is introduced in Chapter 6). The line continues until the patient either experiences the event of interest, or is lost to follow-up. If the line is extended from the circle at the point of diagnosis back until the patient's birth (that is, Age=0), the corresponding value on the x-axis relates to the value of birth cohort that is used as part of the age-period-cohort modelling framework. With the interest in the chapter relating to the modelling of incidence, the points of interest in Figure 2.1 are the circles indicating the point of diagnosis. In order to model this effectively, these points are often grouped together into n-year categories for both age and period. If these categories are too wide, such as 5 or 10 year categories, important changes in the incidence rate may be missed. However, if too fine categories are used, it is argued that this could lead to excessively noisy estimates. A balance could be reached by finely splitting the data into appropriate subsets of the Lexis diagram and employing a smoothing technique to counteract the consequently noisier estimates.

2.4.1. Yearly Intervals & Triangular Subsets

Figure 2.2 gives a second example of a Lexis diagram and shows the appropriate averages necessary when splitting the data finely as proposed by Carstensen [2007].

Carstensen highlights that the necessary values when the diagram is split into triangular subsets for yearly intervals of age and period are different to the conventional averages by a

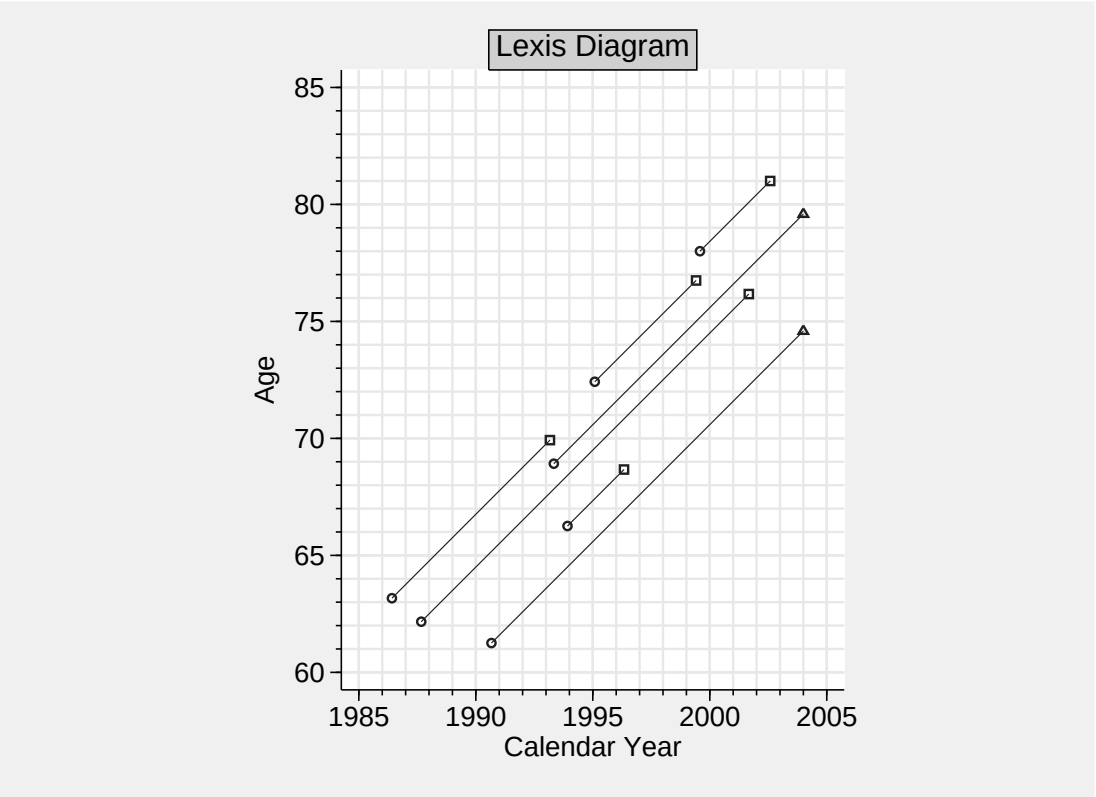

FIGURE 2.1. An example of a Lexis diagram for a small subset of patients. A circle indicates the point of diagnosis, a square indicates a date of death, and a triangle indicates a point at which follow-up ends.

sixth. The conventional averages that would be used for yearly-split data are at the centre of the square; that is, at age $34\frac{1}{2}$ and period $2000\frac{1}{2}$. These values are different from those at the centre of the two triangles by a sixth. Making this distinction provides data that can then model the full extent of the Lexis diagram taking into account both the upper and lower triangular subsets. The reasoning behind the averages used for the values of age, period and cohort is illustrated in Figure 2.2. The set of three lines that pass through the centre of each of the triangles indicates the values that should be used for age, period and cohort. The distinction between the upper and lower triangular subsets is defined by the patients' year of birth. This can again be seen from the partial Lexis diagram given as Figure 2.2; the upper triangular subset relates to patients that are born in 1965, whereas the lower triangular subset relates to patients that were born in 1966. It is also necessary to obtain the appropriate risktime in each of the triangular subsets of the Lexis diagram. Formulae to appropriately calculate these values

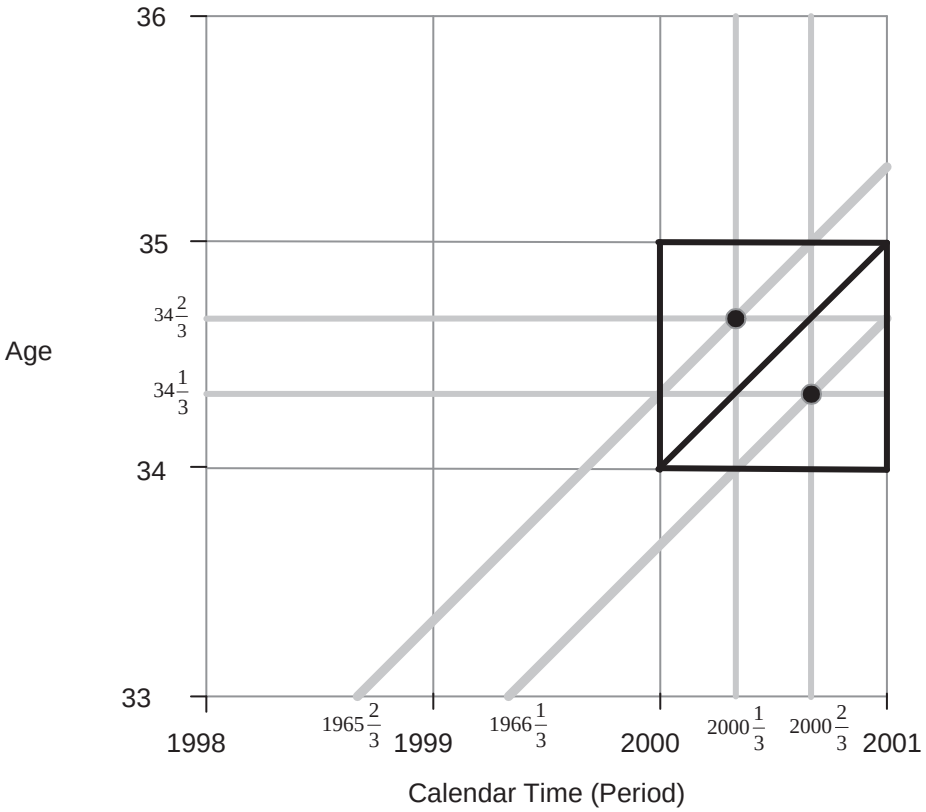

FIGURE 2.2. Snapshot of a Lexis diagram indicating the reasoning behind the use of the average values that are offset by $\frac{1}{6}$ for the triangular subsets (compared to the average values for the squares of a Lexis diagram).

have been proposed [Sverdrup, 1967; Carstensen, 2007] and will be used for the analyses split into triangular subsets. Let $L(a, p)$ be the population size at age a, period p. The risktime in the upper triangular subset (those patients born in the earlier year (cohort $(p-a-1)$) but are diagnosed at age a, and period p) should be calculated as:

$$Y_{a,p,p-a-1} = \frac{1}{3}L_{a,p} + \frac{1}{6}L_{a+1,p+1}, \tag{2.2}$$

whereas in the lower triangular subset (those patients born a year later but are diagnosed at the same age, and period), the risktime is calculated as:

$$Y_{a,p,p-a} = \frac{1}{6}L_{a-1,p} + \frac{1}{3}L_{a,p+1}. \tag{2.3}$$

Although it is common to have the same length of interval for age and period, there has been recent work considering non-equal intervals [Holford, 2006]. Provided that the appropriate averages are used for the subsets of the Lexis diagram, the modelling framework that is described in this chapter can be used for intervals of different lengths. However, it is vital that the risktime is calculated appropriately for the subsets of the Lexis diagram.

2.5. Poisson Models for Rates

Fitting age-period-cohort models utilises the fact that Poisson models can be fitted for rates [Breslow and Day, 1975; Holford, 1980; Frome, 1983]. Further to this, by collapsing over unique covariate patterns, the size of the dataset required to fit the models can be substantially reduced. The data that is required for each observation is the total number of cases in the subset, D, and the total number of population risktime in the subset, Y. This again makes use of the fact that the Poisson likelihood is similar in structure to that required to model rates [Breslow and Day, 1975]. The rate is assumed to be constant within each triangular subset of the Lexis diagram (see previous section). The likelihood contribution for each subset with observation (D, Y) and a constant rate, λ, is:

$$l(\lambda|D, Y) = D\log(\lambda) - \lambda Y. \tag{2.4}$$

This is equivalent to a Poisson log-likelihood for a random variable D with mean λY aside from a constant term. This means that the Poisson distribution can be used in order to model rates provided that an appropriate offset term is employed.

Therefore, a Generalised Linear Modelling (GLM) framework can be used with a Poisson error structure in order to fit the models for rates. The only requirement is that an appropriate offset term of $\log(Y)$ is used when modelling the number of cases D in a Poisson model.

The model that is fitted using the Poisson equivalence to the log-likelihood is essentially:

$$\ln\{D\} = f(\mathbf{a}) + g(\mathbf{p}) + h(\mathbf{c}) + \ln(Y). \tag{2.5}$$

2.6. The Identifiability Issue

The identifiability issue that is associated with APC analyses stems from the fact that the three main terms of the Age-Period-Cohort model form an exact linear relationship. That is:

$$\text{Age} = \text{Period} - \text{Cohort}. \tag{2.6}$$

This linear dependency between the three variables means that unique solutions cannot be obtained when modelling the three variables simultaneously without further constraints. If the modelling of the variables is thought of in terms of matrix transformations, this linear dependency leads to a matrix that is not of full rank. A variety of constraints can be fitted in order to fix the solution for an APC model. However, for the solution obtained to be of relevance, the constraints that are made must be both realistic and sensible.

Figure 2.3 summarises the identifiability issue in the context of incidence; if the values for two of the three variables are known, then it is possible to deduce the value for the third.

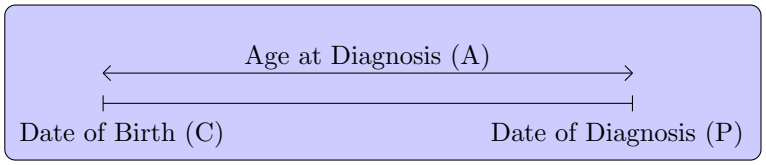

FIGURE 2.3. Illustration of the identifiability issue.

One solution to the identifiability issue is to exclude one of the three variables from the model; this approach has been used in practice [Xie et al., 2011]. However, in doing this, an assumption about the lack of effect of the third variable is being made. To reach a solution to the modelling issues that are proposed by the APC model, it is likely that a strong assumption will be necessary. The trends of disease over time in terms of each of the three variables are of interest to health officials and epidemiologists.

The age variable can provide information on the trends of disease for different age ranges, which obviously impact on the health procedures in place. Uncovering the trends of incidence over the values of age provides key details into the levels of disease to expect in the future, especially with the prospect of an ageing population. The period variable provides an insight into the trends of disease by, for example, being able to highlight the introduction of treatment or screening programs. The trends for the period variable can indicate any major changes in the incidence/mortality rates that effect all age-ranges simultaneously. Finally, the cohort term can be used to assess long-term exposures, with different generations being exposed to different risks [Robertson et al., 1999]. For example, the cohort term could be used to highlight the effect

that malnutrition due to a period of war could have on the levels of disease. An example of this has been seen for colorectal cancer in the Nordic countries and Estonia [Svensson et al., 2005]. With each of the three variables providing different insights into the shape of the disease, it is often not possible to consider the exclusion of one of the variables for the models. At the very least, it is desirable to adjust the two other terms for the effects of the third. Therefore, a number of methods have been suggested in order to attempt to fit appropriate constraints to overcome the identifiability issue in a sensible way.

As has been highlighted in the literature [Holford, 1983; Clayton and Schifflers, 1987b], some of the measures that can be calculated from the age-period-cohort model are identifiable. The linear drift (the combined overall linear trend of period and cohort) can be summarised, and is independent of the parameterisation. Also, the 2^{nd} order differences are identifiable quantities. The 2^{nd} derivatives of the effects have been presented [Clayton and Schifflers, 1987b; McKenzie, 2006]. However, these values can be difficult to interpret.

2.7. Methods for Overcoming the Identifiability Issue

A wide variety of methods have been proposed for overcoming the identifiability issue that is associated with APC models. The vast majority of these methods treat the variables as factors. These methods range from graphical displays of the data to modified modelling techniques in an attempt to avoid the problem that the linear dependency between the variables causes. Methods for overcoming the associated identifiability issue for age-period-cohort models have been given consideration for many years across a number of disciplines. The fact that the effects cannot be fully extracted separately means that thought has also been given to the most appropriate methods for reporting the resulting output from the models.

In practice, simple assumptions are often made in order to achieve an identifiable model. Firstly, the effects of age, period and cohort are often fitted as factors. This results in no functional form being required for the effects of age, period and cohort, but does not circumvent the need to fit further constraints on the model in order to provide a unique solution. Methods for overcoming the identifiability issue traditionally involve restricting the parameters for one or more of the variables in order to fit the model.

The early literature in the social sciences [Greenberg et al., 1950; Mason et al., 1973; Glenn, 1976; Fienberg and Mason, 1979] produced a lot of interest and discussion about the validity of

the assumptions made through the chosen constraints. The lack of identifiability leads to only certain parameters being fully identifiable, and the other parameters can be interchanged and altered depending on the parameterisation that is chosen.

Further study of the lack of identifiability and the methods for overcoming the issue where given throughout the 1980s [Osmond and Gardner, 1982; Holford, 1983, 1985; Clayton and Schifflers, 1987a,b]. The approaches set out by Holford and, Clayton and Schifflers have become the basis for recent methodology in the age-period-cohort setting. The methodological approaches outlined by Heuer [Heuer, 1997] and Carstensen [Carstensen, 2007], that will be investigated in the next section depend heavily on those earlier developments. This approach has also been adopted in practical applications of age-period-cohort analyses [Coviello et al., 2010; Wessler et al., 2010].

More recent methods for fitting age-period-cohort models have adopted more complex methods to fit the underlying model. A lot of the literature has concentrated on Bayesian methodology, [Nakamura, 1986; Berzuini and Clayton, 1994; Besag et al., 1995] whereas the classical statistical literature has adopted approaches using splines in favour of the traditional factor approaches [Heuer, 1997; Carstensen, 2007]. These differences are not only in the methods that are adopted to fit the model but also differ in the approach to the identifiability issue. The new methods largely adopt the residual approach proposed by Clayton and Schifflers [Clayton and Schifflers, 1987b].

There have been other recent proposals with various ways suggested to apply the relevant smoothing, and identification. Methods using generalised additive models have been proposed [Clements et al., 2005], although this approach is proposed in the setting of making projections (see Chapter 4). Also a method has been proposed to estimate a canonical parameter that includes the 2^{nd} order differences and is expressed in terms of acceleration [Kuang et al., 2008]. However, this approach does not lead to simple interpretable estimates.

An approach using partial least squares regression has been outlined [Tu et al., 2011] and applied to Scottish blood pressure data. The approach suggested is similar to a principle components analysis and the analysis can be carried out in spite of the identifiability issue. However, the resulting output is again difficult to interpret.

An intrinsic estimator approach has also been proposed as an alternative method of analysis for age-period-cohort data [Yang et al., 2004]. This approach uses the Moore-Penrose inverse of the design matrix in order to extract an indentifiable solution. This work has been applied and compared to a more standard constraint approach via simulation [O'Brien, 2010]. The results of the simulation highlight that consistent estimates are only observed when similar constraints are applied for the analysis to those applied in the data generating process.

A number of further comparisons of various approaches have been made in various settings [McNally et al., 1997; Robertson et al., 1999]. McNally *et al.* [1997] compare the Clayton and Schifflers approach [Clayton and Schifflers, 1987b] (extracting the linear drift) to two further approaches that attempt to further distinguish between the linear period and linear cohort effects [Robertson and Boyle, 1986; De Carli and La Vecchia, 1987]. They conclude that it is controversial to try to overcome the identifiability issue any further than by extracting the identifiable quantities and that the approaches that try to do so should not be over-interpreted. Robertson *et al.* [1999] draw similar conclusions when making a comparison of available approaches to overcoming the identifiability issue in a GLM framework.

The majority of the methods that have been proposed to overcome the identifiability problem apply arbitrary constraints to the model in order to give separate age, period and cohort effects. Also, the approaches generally use data that is aggregated into large age and period categories (typically 5-years) and fitted in a factor model framework. A recent paper by Rosenberg and Anderson [2011] argues that the age-period-cohort approach has been underused. They argue that the issue of indentifiability and the lack of understanding around the models have limited their use in practice. They argue that the parameterisations that extract the drift and present the non-linear components of period and cohort give the potential to obtain identifiable quantities that are useful for epedimiologists. In the following section, an approach is detailed that estimates the identifiable quantities whilst also smoothing the effects of age, period and cohort using splines.

2.8. APC Analysis using Restricted Cubic Splines

The method advocated by Carstensen using restricted cubic splines [Carstensen, 2007] provides an intuitive solution to the identifiability issue whilst allowing a continuous interpretation of the variables. The use of splines for age-period-cohort models was first proposed by Heuer

[1997]. The method is described in Section 2.8.2, having first introduced the methodology behind restricted cubic splines. An example using Finnish registry data is provided, and the extensions to the method proposed by Carstensen in terms of interaction effects with other key covariates are demonstrated through the example. Applications using a spline approach for age-period-cohort models have recently been appearing in the literature [Sala et al., 2009; Coviello et al., 2010].

2.8.1. Restricted Cubic Splines

A spline is a mathematical function defined by a collection of piecewise polynomials joined at a pre-defined number of points; known as the knots. The first and last of these points are often referred to as the boundary knots. A spline is constrained in order to produce a smooth overall curve. The function that is fitted is forced to have continuous 0^{th}, 1^{st}, and 2^{nd} derivatives; that is, the fitted curve is $\mathcal{C}^2$ continuous. Restricted splines impose the further condition that at, and beyond, the boundary knots, the fitted function is forced to be linear. The selection for the placement and quantity of the knots is usually user-defined and can affect the overall fit of the function. This is a criticism that has been levelled at the use of splines. However, in a wide-ranging set of applications, there have been numerous sensitivity analyses to assess the effect of knot selection [Harrell et al., 1988; Heinzl et al., 1996; Lambert et al., 2010]. These analyses have, on the whole, shown that provided a sufficient number of knots are used, knot selection does not affect the outcome of the analysis too heavily.

Restricted cubic splines refer to restricted splines that use cubic polynomials between the knots. It has been shown that restricted cubic splines can be used in many forms of regression analysis [Durrelman and Simon, 1989]. Restricted cubic splines have been used an a wide variety of settings for applied research, from modelling dose-response in logistic regression [Muscat et al., 2000] to flexibly capturing the shape of the baseline hazard in survival analysis [Royston and Parmar, 2002]. The approach set out by Royston and Parmar will be considered further in Chapter 6.

A restricted cubic spline function can be written in terms of the $K-1$ basis functions, $B_i(x)$. For knots $k_1, \ldots, k_K$, the spline function, $S(x)$, can be written as:

$$S(x) = \gamma_0 + \sum_{i=1}^{K-1} \gamma_i B_i(x), \tag{2.7}$$

where

$$B_1(x) = x \tag{2.8}$$

and, for $i = 2, \ldots, K-1$

$$B_i(x) = (x - k_i)_+^3 - \lambda_i(x - k_1)_+^3 - (1 - \lambda_i)(x - k_K)_+^3 \tag{2.9}$$

where $(x - k_i)_+^3$ is equal to $(x - k_i)^3$ if the value is positive and 0 otherwise. The λ values are defined as:

$$\lambda_i = \frac{k_K - k_i}{k_K - k_1}. \tag{2.10}$$

Due to the potential for high levels of correlation between the basis vectors, it is recommended that the vectors are orthoganalised when using the spline basis in a regression setting. In the analyses carried out using restricted cubic splines in this chapter, Gram-Schmidt orthogonalisation [Golub and van Loan, 1996] has been applied. The basis vectors form the design matrix that can be used in the regression approach suggested by Carstensen [Carstensen, 2007]. The use of splines in age-period-cohort modelling had also been suggested 10 years earlier by Heuer [Heuer, 1997]. Both of these authors discuss the decision of the choice of knots for each of the three variables; age, period and cohort. In the following chapter, a detailed analysis of the impact of knot selection and the placement of the knots will be given.

2.8.2. Using Splines for the APC Analysis

In essence, the method proposed by Carstensen uses restricted cubic (natural) splines for the age, period and cohort terms within a Generalised Linear Model (GLM) framework with a Poisson family error structure, a log link function and an offset of log(person risk-time). However, in order to overcome the identifiability problem, transformations are made to the spline basis vectors for the period and cohort effects using matrix transformations. These matrix transformations to the spline design matrices are used to remove the trend from the period and cohort matrices by projecting onto the linear space.

The transformations are made to the spline basis vectors so that the resulting output has a clear and sensible interpretation in spite of the identifiability issue. The matrix transformations are performed to remove the linear trend from the Cohort and Period terms. The so-called

drift term is then added to either of the Cohort or Period terms, depending on the selected parameterisation.

The appropriate spline basis vectors are combined into matrices relating to each of the components of the model (Age, Period and Cohort). Let these three design matrices be $\mathbf{M}_A$, $\mathbf{M}_P$ and $\mathbf{M}_C$. The method requires that the Period and Cohort matrices be detrended. This is achieved by projecting the columns of the matrices onto the orthogonal complement of a two column matrix, $\mathbf{X}$. In the case of the detrending of the Period matrix: $\mathbf{X} = [1 \mid P]$, where P is the column of all of the values of Period.

The form of a general inner product that allows weighting is:

$$\langle \mathbf{x} \mid \mathbf{y} \rangle = \sum_i x_i w_i y_i = \mathbf{x}'\mathbf{W}\mathbf{y}, \tag{2.11}$$

where $\mathbf{W} = \text{diag}(w_i)$. The projection matrix on the column space of $\mathbf{X}$ with respect to a general inner product is:

$$\mathbf{P}_\mathbf{W} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}, \tag{2.12}$$

and the projection of $\mathbf{M}$ on the orthogonal complement is:

$$(\mathbf{I} - \mathbf{P}_\mathbf{W})\mathbf{M}. \tag{2.13}$$

Once the Period and Cohort matrices have been projected in this way, the next stage is to reduce the number of columns of the matrices in order to ensure that they are of full rank. The columns of the matrices are also pivotted during this process. The rank of the matrices, and the pivotting vector to be used, are obtained by using QR decomposition with column pivotting [Golub and van Loan, 1996]. The columns required to ensure the matrices are of full rank are then selected. The matrices are then centred around the relevant reference points by subtracting a row corresponding to the reference point from each of the rows of $\mathbf{M}$. A column of 1s is attached at the beginning of the $\mathbf{M}_A$ matrix in order to ensure that the intercept is part of the Age effects. The intercept term is contained within the Age effects so that the Age term carries the rate dimension. Then, according to the parameterisation, the drift column is added to the front of either the Period or Cohort matrix. Full details of the matrix operations are given in the Appendix of Carstensen's paper [Carstensen, 2007].

The weighting matrix used for the projection can take on any form, however, three logical choices for the weights are: $w_i = 1$, $w_i = D_i$ (where D_i is the number of cases for an observation), and $w_i = Y_i$ (where Y_i is the population risk-time for an observation). Carstensen suggests using a weighting that is based upon the number of cases (D) [Carstensen, 2007]. Using equal weights (of 1) during the process of the detrending is a method that is attributed to Holford [Holford, 1983]. Using different values for the weights produces different estimates for the drift term.

Having performed the matrix transformations, a GLM is fitted using the adjusted spline basis vectors. Using this GLM as a foundation, it is possible to extend the analysis to include covariates and also interaction terms between covariates and the key variables of interest. Statistical testing can be used to determine if the effect of a covariate is significantly different across the range of the timescales (Age, and Period).

2.8.3. apcfit

As part of the PhD thesis, software has been developed in order to fit these models in the statistical software package; Stata [StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP]. The code written to carry out the command is given in Appendix I. The corresponding article describing the command is given in the Stata Journal [Rutherford et al., 2010] and a copy is given in Appendix II. The developments of the method to incorporate a "reduced" set of splines for interaction effects with other key covariates are introduced in the paper, and illustrated through an example.

The command to carry out the model fitting is called `apcfit`, and is used to carry out the matrix calculations detailed in the previous section and to fit the consequent GLM. A second command that is available as part of the user-written Stata package is `poprisktime` which calculates the appropriate population risktime for the triangular subsets of the Lexis diagram using the formulae detailed in Section 2.4.1. The code written to carry out `poprisktime` is given at the end of Appendix I.

The `apcfit` command has been modified and further developed to carry out incidence projections. The command is also used for the methods using splines that are detailed in Chapter 4.

2.8.4. Interpreting the Effects

The interpretation of the fitted values for the three variables depends on the parameterisation that is adopted. In light of the identifiability issue, it is clear that firm conclusions surrounding the three effects should be made cautiously. A thorough investigation of the various models and parameterisations can give only a guideline to the underlying true effects. The arbitrary allocation of the drift term that is a feature of these models is indicative of the danger of over-interpreting the absolute values for the Rate Ratios (RRs). However, the change in gradient, or curvature, of these curves is an identifiable quantity. The key concepts in interpreting the effects given in these models are shown through an example in the following section.

2.9. Example

A colon cancer dataset from Finland will be used to illustrate the use of the method. The analyses will be carried out using a user-written command in Stata [Rutherford et al., 2010]. The data cover diagnoses between 1980 and 2004 for all regions of Finland. It was decided to restrict the age range of the dataset to people less than 80, and greater than or equal to 20 years of age at the date of diagnosis. In order to highlight the possibility of including covariates into the analysis, the gender of patients was included when collapsing the dataset into unique records of age, period and cohort. The data were collapsed into yearly intervals for age and period, leading to an upper and lower cohort value for each unique combination of age and period (according to the patient's date of birth). These relate the triangular subsets of the Lexis diagram which were highlighted in Figure 2.2.

This led to $(80-20)\times(2004-1980)$ different age-period categories, each of which were subdivided by date of birth into two further categories. This gave a total of 2880 observations for (D,Y); one for each triangular subset. However, as the Finnish dataset contains sufficient information to include a sex term into the dataset, the dataset actually contains 5760 ($=2880\times 2$) observations. In order to increase the dataset to include the sex term, the calculations for population risk-time (detailed in Section 2.4.1) were performed for each of the genders.

There are 10 key interpretation points that will be highlighted throughout the example. The interpretation points are numbered and given in italics.

Firstly, key plots to summarise the rates over the variables are given in Figure 2.4. These plots should always be conducted prior to an age-period-cohort analysis as these figures show

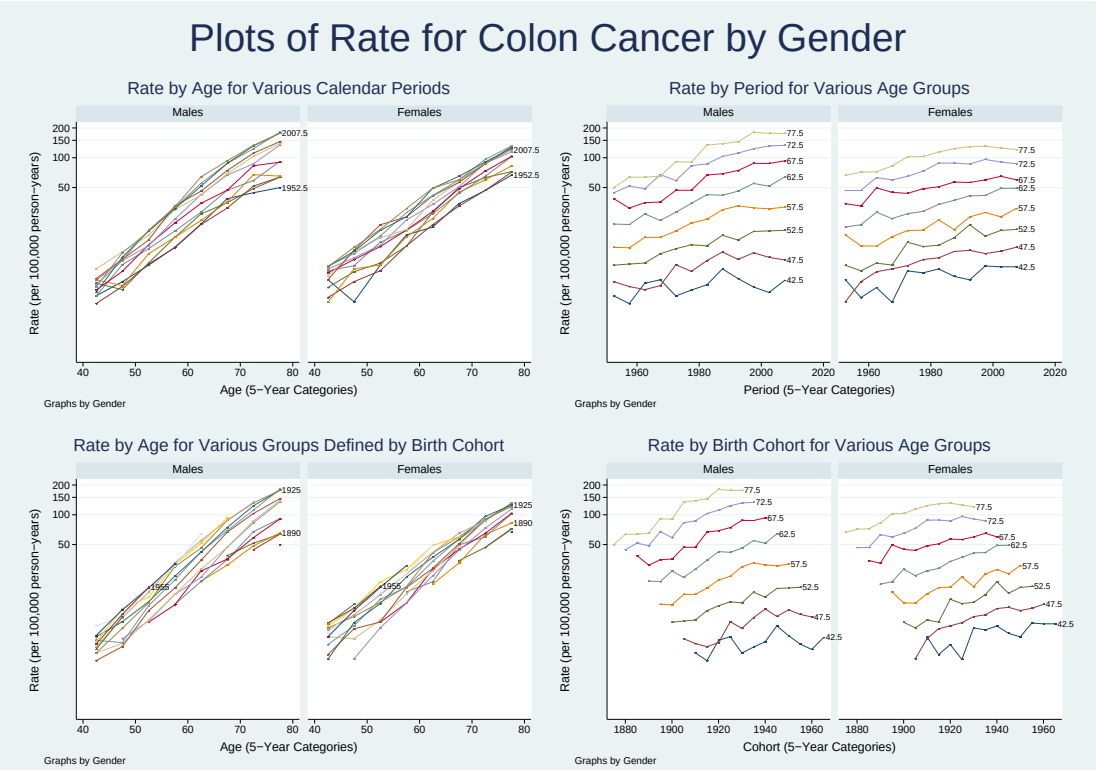

FIGURE 2.4. Combined plot of the summary of the rates by the key variables for colon cancer patients in Finland.

how the observed rates vary over the values of the key variables of interest. They can be used to assess whether the age-specific rates are approximately proportional between periods and cohorts respectively. In order to construct these plots, the data needs to be tabulated into more traditional 5-year windows for age, and period to smooth the resulting plots. However, the more finely split data can be used for the spline models as the splines can be used as a smoothing tool.

It is clear from Figure 2.4 that the effect of age is fairly linear (on the log scale) for both genders. The plots also illustrate the effect of period and cohort, and are useful in illustrating which age-categories are used to define the estimates for the effect of cohort at any particular point. That is, that the effect of cohort in the 1960s can only be estimated by the very youngest patients under study, whereas the effect of cohort in 1880 is defined solely by the experience of the oldest patients under study.

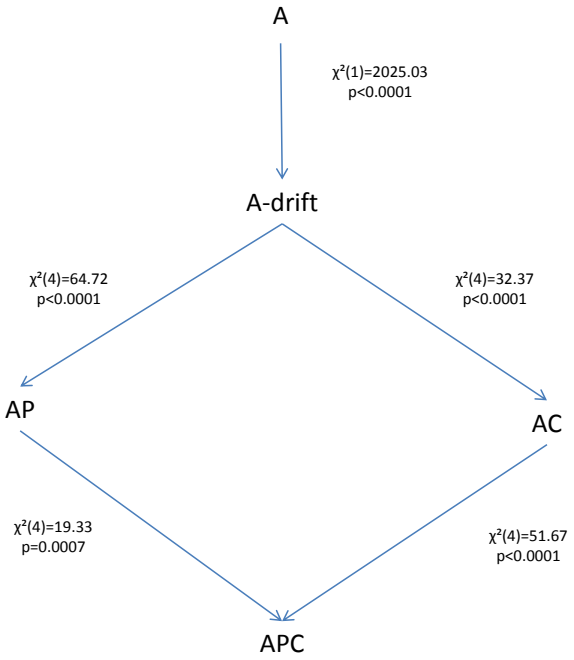

FIGURE 2.5. Flow chart highlighting the likelihood ratio tests between the various nested models.

As detailed in the previous section, restricted cubic spline functions are used as the functional form for the 3 functions detailed in Equation 2.1. The spline functions shall be denoted with an extra S for each of the functions. A choice is required for the number of knots (degrees of freedom) for each of the three spline functions. For this initial exposition of the approach using splines, 5 degrees of freedom are used for each of the spline terms for age, period and cohort; a detailed analysis of the choice for the number of knots is given in Chapter 3. The full APC model using splines will be denoted as:

$$\ln D = f_{S_5}(\mathbf{a}) + \tilde{g}_{S_5}(\mathbf{p}) + \tilde{h}_{S_5}(\mathbf{c}) + \delta_{p/c} + \ln Y, \tag{2.14}$$

where, for example, f_{S_5} denotes the spline (S) function for age, with 5 degrees of freedom. The $\tilde{g}$ indicates the constraining of the function in order to extract the linear drift term, represented by δ. The subscript p or c is used to indicate which of the terms has been allocated

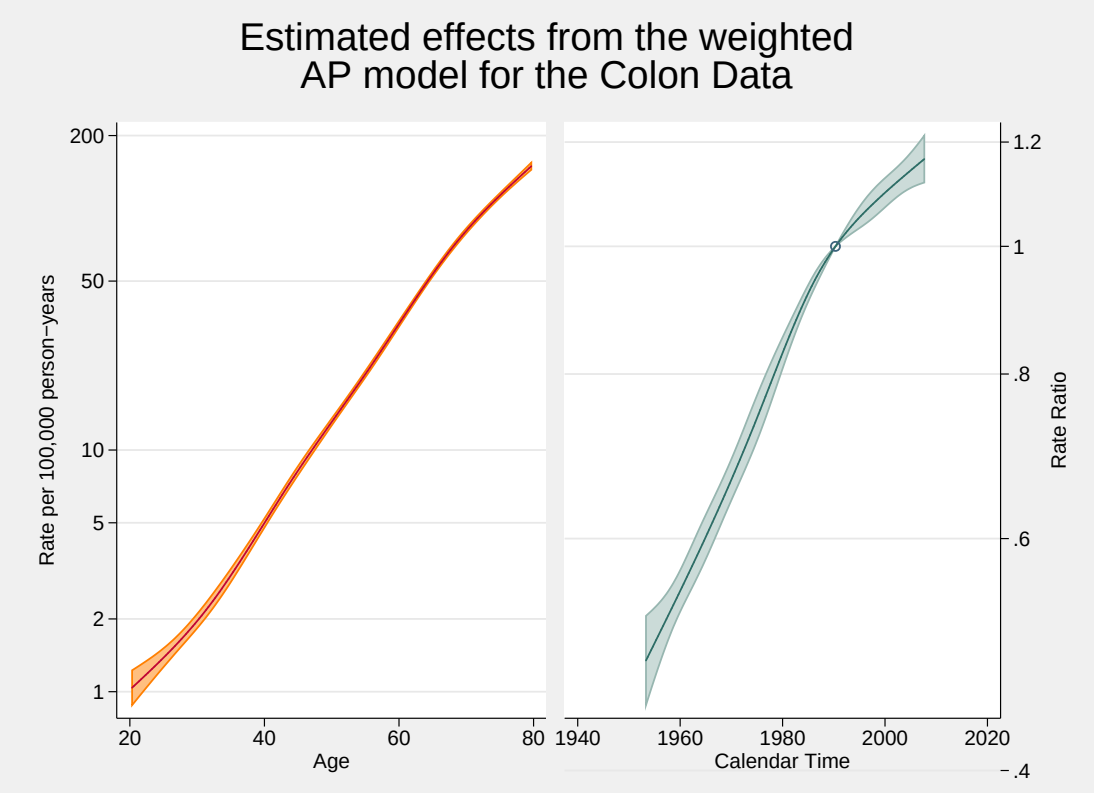

FIGURE 2.6. Fitted values for the age-period model for the colon cancer data.

the drift. The knot placement for the 4 internal knots are decided by the centiles of the relevant variables; the knots are placed equally across the centiles. Alternative knot placements will be considered in Chapter 3.

Figure 2.5 shows the hierarchical structure of the set of models that build towards the full age-period-cohort model (desribed in Equation 2.14). The arrows indicate the models that are nested within each other and, therefore, can be tested with the Likelihood Ratio Test (LRT). The p-values indicate that the full age-period-cohort is required to best capture the information.

Figure 2.6 shows the fit if only the age-period model is fitted. This can be described in a similar form to that expressed for the full age-period-cohort model given in Equation 2.14 by:

$$\ln D = f_{S_5}(\mathbf{a}) + g_{S_5}(\mathbf{p}) + \ln Y, \tag{2.15}$$

The linear drift term is a component of this model (contained within the unconstrained $g_{S_5}(\mathbf{p})$ term), as is the case if the age-cohort model is fitted (see Figure 2.7). The function $g_{S_5}(\mathbf{p})$ is

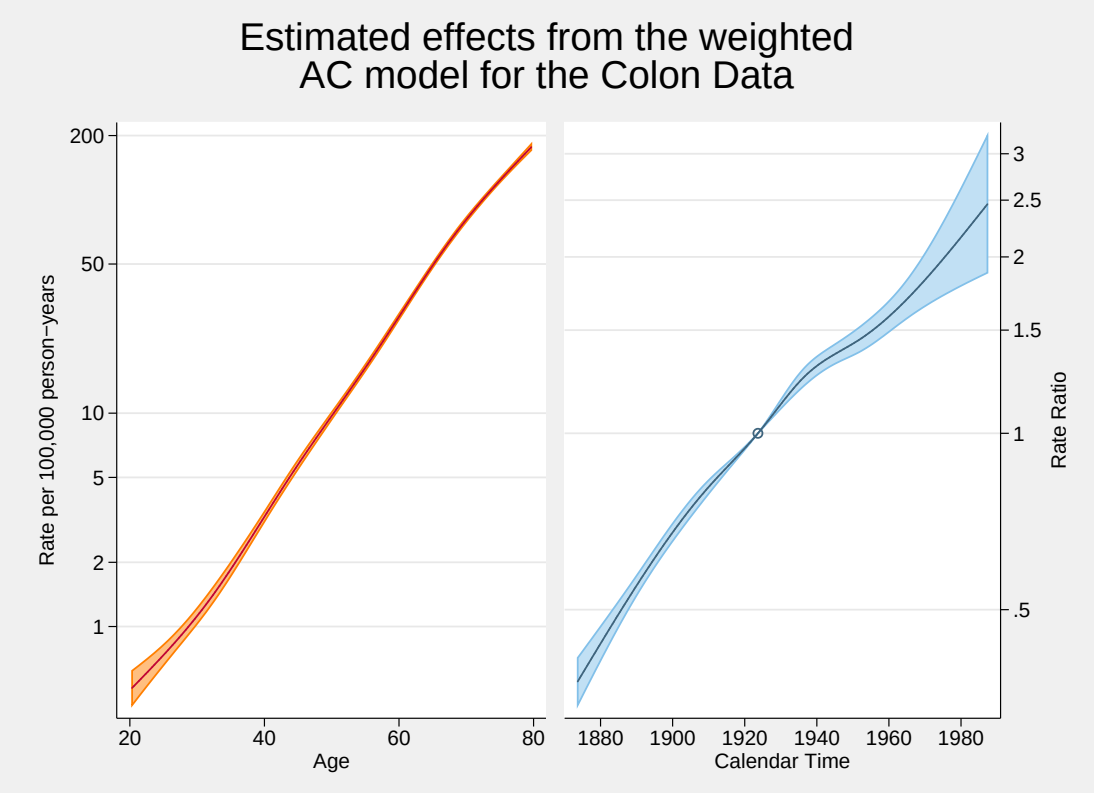

FIGURE 2.7. Fitted values for the age-cohort model for the colon cancer data.

zero for the reference period, p_0. When fitting the AP model, an assumption about the lack of effect for the cohort term is made. The likelihood ratio statistics given in Figure 2.5 indicate that the additional effect of the cohort term is highly statistically significant at the 5% level. This model assumes that the effect of period is proportional across the age range. The left-hand side of Figure 2.6 gives the effect of age on the rate scale in the reference period (p_0=1986.67). In 1987, approximately 1 per 100,000 patients aged 20 are diagnosed with a case of colon cancer each year in Finland. It is clear that there is a strong effect of age as by age 80 the effect is closer to 150 cases per 100,000 person-years. The right-hand side of the figure indicates the effect of period as a rate ratio. The values can be considered as incidence rate ratios compared to the reference period. The incidence of colon cancer has been increasing over calendar time.

Figure 2.7 shows the graph for the age-cohort model. This model is formulated in the same way as in Equation 2.15 with the g function replaced by h. Similarly to the conclusions drawn around the age-period model, it is clear from the LRT that this model is not adequate compared

to the full APC model. The interpretation of this figure is similar to that of Figure 2.6. The age curve is given on the rate scale and shows the incidence rate over age in the reference cohort, c_0 (=1926.33). The cohort curve shows the incidence rate ratio for year of birth relative to the reference cohort. These two graphs are included for completeness, but in this example it is clear from the results of the likelihood ratios tests that the full APC model is required.

The first two key interpretation points are:

(1) *The AP and AC models do not suffer from an identifiability issue. However, it is important to remember that excluding one of the terms (either Period or Cohort) is equivalent to assuming that it has no influence on the Age and the other remaining term. This, clearly, is a strong assumption especially in the case when the APC model is statistically significantly better than the AP or AC models. These comparisons can be made by using Likelihood tests to compare the nested models.*

(2) *Differences between an AP model and APC model are due to an "adjustment" for Cohort. That is, the effect of period is allowed to be different according to the cohort of birth. This can be thought of as similar in principle to an interaction. The effect of period is no longer proportional across the entire age-range as it also depends on the year of birth for a given patient. This shows the interplay between the two timescales; age, and calendar time.*

Figure 2.8 shows the results of fitting the full APC model with an equally-spaced (uniform) knot placement, the drift term allocated to the cohort variable, and 5 degrees of freedom for each of the spline functions. It is clear from the graph that the period term is constrained to be 1 on average (technically, 0 on average on the log-scale). The linear drift, which is a combination of the linear coefficients for both the period and cohort terms is entirely allocated to the cohort term in this parameterisation. The drift is an identifiable quantity in that it does not change for any given parameterisation of the full APC model. In Figure 2.8 the age effects are given as the effects of age in the reference cohort. The reference cohort is indicated by the circle on the Cohort term effects, and is 1926.33. The period and cohort terms act as incidence rate ratios. The three terms need to be multiplied in order to obtain the incidence for any given age and period combination.

Four further key interpretation points are:

(3) *If only one of the reference points is fixed of the period or cohort terms then the age-effects that are given are log age-specific rates in that reference period (p_0) or cohort (c_0) after adjustment for the other variable.*

(4) *The drift term is simply the combined slope for the period and cohort terms. It can be allocated to either period and cohort dependent on the parameterisation. This allocation is somewhat arbitrary and it very much depends as to what is desired from the model as to which parameterisation should be selected.*

(5) *The overall drift that is included into one of the period or cohort terms, is an overall value to describe the linear trend. It does not give a true indication of times when the linear trend is above or below this averaged overall value for the combined linear trend.*

(6) *The interpretation of the age, period and cohort terms is, in some ways, dictated by the "wiggliness" of the fitted lines that can be plotted. This "wiggliness" is inherently associated with the degrees of freedom that are used for the spline terms. Therefore, it is important to select the number of knots (degrees of freedom) and the placement of the knots carefully. If a large value for the degrees of freedom is selected then there is the distinct possibility that random noise will be picked up by the fit of the splines. However, if too few knots are selected then there is the potential to over-simplify the underlying shape for each of the variables. This may lead to important changes in the levels of incidence/mortality being overlooked when trying to make statements about the trends of a disease over time. It is important to try to quantify and understand the reasons for the deviations from the linear trend of the disease. These deviations could be used to indicate changes in practice, or changes in exposure for different generations or for the population as a whole, depending in which variable the deviations are manifested.*

In the model fitted to produce Figure 2.8 there are a potential 166 unique Cohort values to choose from to be the reference Cohort value. The choice of 1926.33 is based on taking the median Cohort value, using a frequency weight based on the number of cases.

Figure 2.9 shows the fitted values for age and cohort for 3 different reference cohorts. The line for the fitted age values for the reference cohorts of 1926.33 and 1980 are approximately 0.69 apart on the log scale. The value for exp(0.69) is approximately 2. It can be seen from

Figure 2.8 that the RR for a value of 1980 is almost twice that of the value for the 1926.33 cohort value. This shows how altering the reference cohort affects the interpretation of the age effects. The shape of the curves for age and cohort are equivalent and the curves are just shifted up and down depending on the chosen reference value.

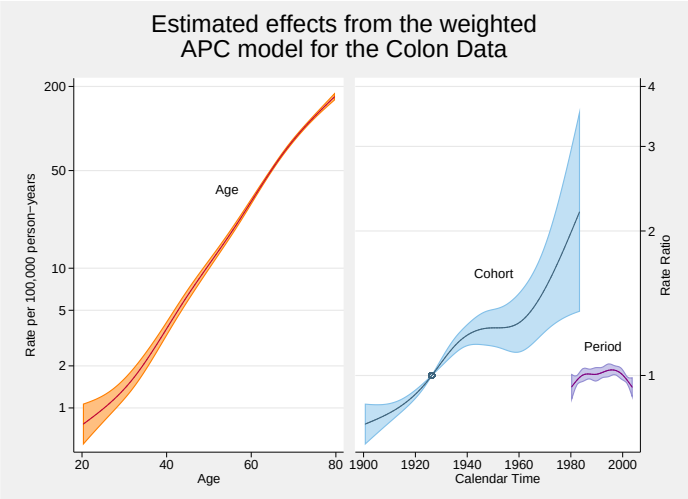

FIGURE 2.8. Fitted values for the age-period-cohort model for the colon cancer data. The drift is allocated to the cohort term, and the age effects give the rate in the reference cohort.

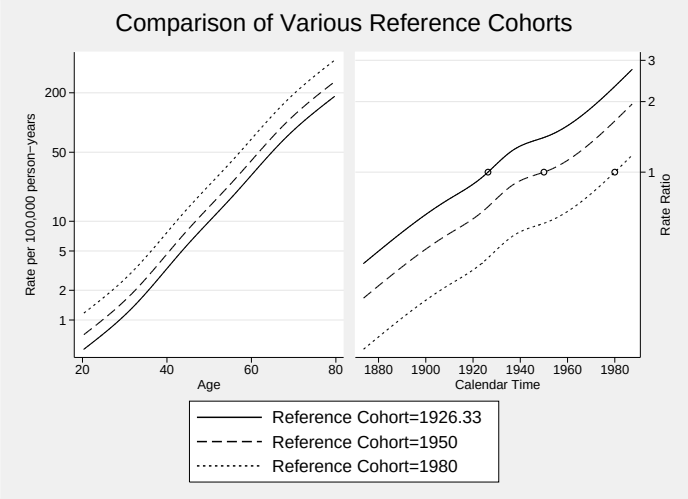

FIGURE 2.9. Comparison of the age effect (on the rate scale) for various reference cohorts.

The linear drift is combined with the Cohort effects in the parameterisation that is shown in Figure 2.8. The linear drift is a combination of the slopes evident from the data for both the

period and cohort terms. Therefore, the values for the rate ratios given for period and cohort in Figure 2.8 should not be over-interpreted. However, the deviations from the linear effect (the curvature of the fitted line) can be fully interpreted. Potentially, these deviations can be more apparent in the constrained term as opposed to the term with the allocated drift.

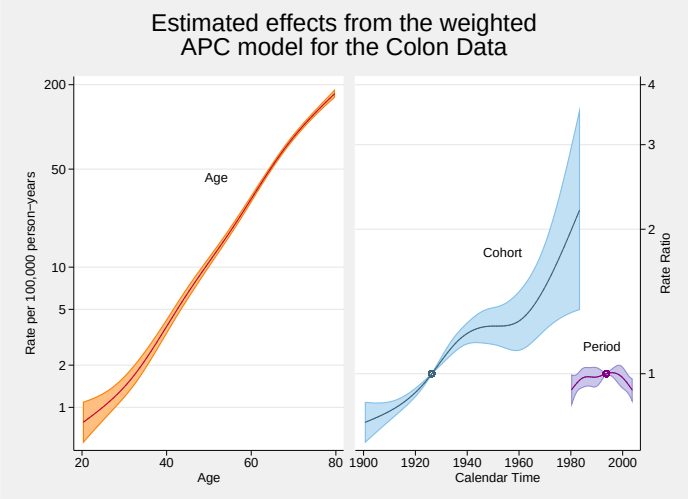

FIGURE 2.10. Fitted values for the age-period-cohort model for the colon cancer data. The drift is allocated to the cohort term. Both a reference period and reference cohort are fitted, which alters the interpretation of the age effect.

Figure 2.10 fits the same parameterisation as in Figure 2.8 but instead of fixing just a single reference point it fixes a reference point for both Period and Cohort. This additional reference point modifies the interpretation of the Age effects.

(7) *If both of the reference points are fixed, then the age values take on a different interpretation that is related to* $a_0 = c_0 - p_0$. *If both reference points are fixed, then the age effects at* a_0 *equal the fitted rates for period* p_0 *and cohort* c_0. *That is, the age-effects in the cohort* c_0 *are shifted (by multiplication) to be relative also to* p_0 *at* a_0. *The value given for* p_0 *when the period effects are constrained and forced to be zero on average, is the multiple by which the age-effects are adjusted. This adjustment alters the interpretation of the age-effects into a less intuitive form.*

In this example, the Period effect is not altered dramatically by fixing the median Period value as the reference point. The introduction of a reference point for the Period effect simply applies a vertical shift to the Period curve, on the log scale. The factor by which the curve

is moved depends upon the value for the reference point in the curve that is obtained by exponentiating the values that are averaged to be 0 on the log scale. If the value is close to 1 then only a small shift will be apparent.

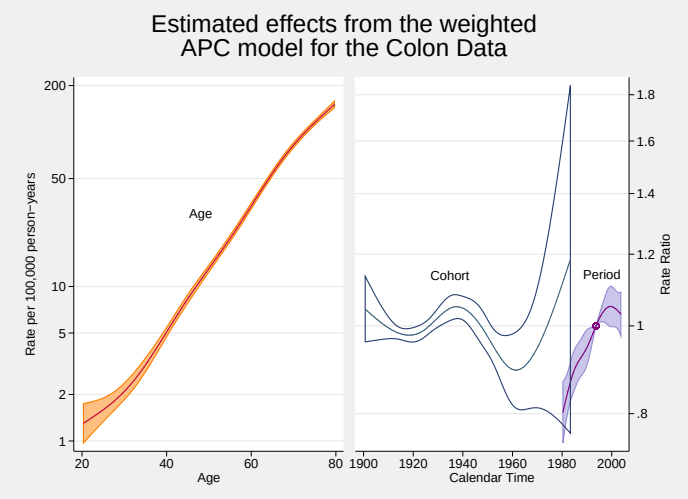

FIGURE 2.11. Fitted values for the age-period-cohort model for the colon cancer data. The drift is allocated to the period term, and the age effects give the rate in the reference period.

Figure 2.11 indicates an alternative parameterisation of the model using the same dataset. In the parameterisation represented in this figure, the linear drift is allocated to the Period variable rather than the Cohort variable. This means that the Cohort fitted values are constrained to not have a slope, and also to be 0 on average on the log scale. The age-effects can now be interpreted as the fitted values for Age in the reference Period (as opposed to the reference Cohort for the previous parameterisation exemplified in Figure 2.8). The modifications that can be made to the age-effects by altering the reference Period are concurrent with the statements made when discussing the reference Cohort in Figure 2.8. Any "dips" in the fitted effects for the Period variable are second-order features of the curve and are therefore not "an artefact of the parameterisation" [Carstensen, 2007; Engholm et al., 2009]. However, they could be an artefact of over-parameterisation by using too high a degrees of freedom for any of the spline terms.

It is also possible to fit a model that entirely excludes the drift from both the Period and Cohort terms. These models involve constraining both the Period and Cohort fitted values to have no overall slope. However, it is still possible to draw information from the constrained

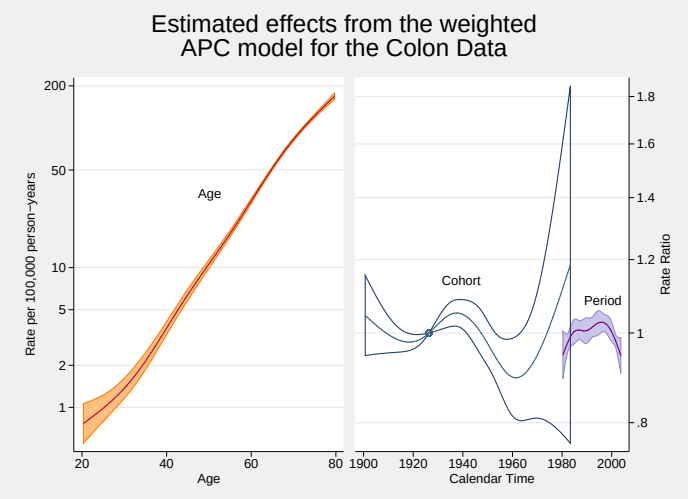

FIGURE 2.12. Fitted values for the age-period-cohort model for the colon cancer data. The drift is not allocated to either term. The period and cohort effects give the non-linear effects and are constrained to be zero on average on the log scale. The age effects are relevant for the reference cohort.

values. The changes in gradient (the second order terms) are consistent across all the parameterisations. Therefore, it is possible to draw conclusions surrounding time points of significant changes in terms of the levels of incidence. Even in the case when the drift is excluded entirely, the fixing of a reference point for either the Period or Cohort term (or both) determines how the Age effects are interpreted. An example of the results that are obtained when the linear drift is entirely excluded is given in Figure 2.12. It should be noted that when the drift is excluded entirely, it is not appropriate to combine the three effects to give the estimates of the overall fitted rates.

(8) *The constrained term without a reference point averages to zero on the log scale. If a reference point is included then the "flat", "slope-less" curve produced for the constrained effect can be pivoted around the reference point to give the unconstrained (that is, when the drift is included with the term) effect. This pivotting gives a nice feature that indicates the form of the drift term and how the drift is allocated. This can be seen by comparing Figures 2.8 and 2.12.*

(9) *The constrained term can still give information about times of change even though it is constrained overall. This is because the* 2^{nd} *order differences are identifiable.*

2.9.1. Including an Interaction Term

The method that has been discussed for fitting an APC model can also be adapted in order to take interaction terms with other covariates into account. The inclusion of further covariates can give further insight into the levels of incidence that are observed. For the Finnish colon cancer data, it was decided to split the data by a covariate for gender in order to compare the levels of incidence for males and females.

The form of the model fitted to generate an age-sex interaction (with 8 degrees of freedom for each of the terms) is:

$$\ln D = f_{S_8}(\mathbf{a}) + \tilde{g}_{S_8}(\mathbf{p}) + \tilde{h}_{S_8}(\mathbf{c}) + \delta_{p/c} + \ln Y + sex + sex * f_{S_8}(\mathbf{a}). \tag{2.16}$$

The use of 8 degrees of freedom for the underlying shape is purely for illustrative purposes to highlight the potential to overfit the interaction. This model will be compared to a second model with a "reduced" set of splines for the interaction term:

$$\ln D = f_{S_8}(\mathbf{a}) + \tilde{g}_{S_8}(\mathbf{p}) + \tilde{h}_{S_8}(\mathbf{c}) + \delta_{p/c} + \ln Y + sex + sex * f'_{S_3}(\mathbf{a}). \tag{2.17}$$

In this model, the underlying age effect still has 8 degrees of freedom but the age-interaction with gender only has 3 degrees of freedom because of the simpler spline function for age ($f'_{S_3}(\mathbf{a})$) that is used for the interaction. The interaction between age and sex allows the effect of sex to vary across the range of values for age. This form of interaction is similar to what is known as a time-dependent effect in survival analysis, which will be covered in Chapter 6.

Figure 2.13 gives an example of the graphical display that can be obtained having included a time-dependent interaction term into an APC model. The figure highlights the potential benefit of using a reduced number of degrees of freedom for the interaction term compared with the number of degrees of freedom used for the underlying variable. Using a large number of degrees of freedom for the interaction can lead to a case of over-fitting. The dashed lines represent the fit for the RR for males to females for the age-dependent effect with an underlying age effect having 8 degrees of freedom and the interaction being allowed 8 degrees of freedom. It is clear that the shape that is produced from having such a large number of degrees of freedom for the interaction is unrealistic. A more realistic approximation to the curve is given by the solid line. The solid line represents the reduced interaction. The solid line is produced from a

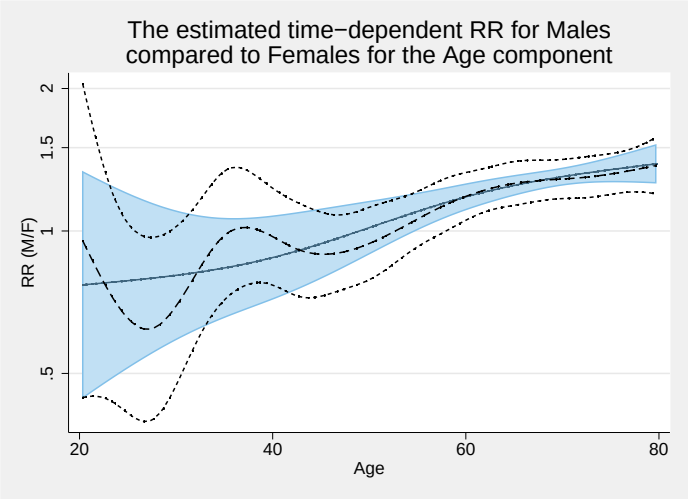

FIGURE 2.13. Graph of time-dependent IRR for the age-by-sex interaction. The solid line is the RR (Males/Females) for the age-by-sex interaction using a reduced set of splines with 3 degrees of freedom (from the model using reduced splines for Age and Cohort having fitted a model with 8 degrees of freedom for each of the components). The shaded region around this line gives the relevant 95% confidence interval. The dashed line gives the IRR for sex in terms of the Age component from the model with the full interaction for sex (from the previous section) using 8 degrees of freedom for each of the components. The dotted lines form the appropriate 95% confidence interval.

model with 8 degrees of freedom for the underlying age effect but has only 3 degrees of freedom for the age-dependent interaction with gender. It is evident from Figure 2.13 that this fitted line provides a more realistic effect for the age-gender interaction. For ages less than 50, males appear to be at a decreased risk of colon cancer for the Finnish dataset. However, after the age of 50 it appears that the risk of incidence is increased for males, compared to the females. The 95% confidence intervals are given by the regions surrounding the fitted lines.

(10) *Interaction terms are possible with these types of model. However, care must be taken to consider the choice of parameterisation when interpreting the interaction effects. Using a "reduced" set of splines may lead to a better fit for the interaction.*

2.10. Discussion

Age-period-cohort models provide a useful modelling tool for assessing the rates of disease over calendar time and age. The issue of the lack of identifiability of the model is one that cannot be solved. However, it is possible to make sensible constraints on the model in order

to extract the identifiable quantities. It is also possible to present the results of the analysis graphically provided that the constraints on the model are fully explained in tandem.

Using restricted cubic splines for the age-period-cohort analyses means that questions will be raised about the selection of the knots. The fact that the number and placement of the knots is user-defined means that it could add further issues for the analyses. An analysis into knot selection for these models will be the topic of the next chapter.

The issue of allocating the drift term depends on the subject matter at hand. However, the drift term still cannot be separated into whether the trend is due to period or cohort effects. If interest lies in the effect of cohort then it makes sense to allocate the drift term to the cohort term when presenting the results. However, the same model is being fitted irrespective of the allocation of the drift term.

Further extensions of the age-period-cohort models are for use in projecting estimates of incidence. Any projection technique should be independent of the chosen parameterisation for the model [Osmond, 1985]. Projecting incidence from age-period-cohort models will be the topic of the fourth chapter.

The extension of using a reduced degrees of freedom for the interaction term appears to be an important contribution. There is a tendency to overfit the interaction terms if using the original main effects spline terms to construct the interaction. The fact that the software development in Stata [Rutherford et al., 2010] allows the user to fit further models with the original spline terms allows further scope to the model-fitting process. The software has been used in the literature [Coviello et al., 2010], and there is an intention for the software to be used in an international comparison of lung cancer incidence by IARC.

CHAPTER 3

How many knots and where to put them? Age-period-cohort models using restricted cubic splines.

3.1. Chapter Outline

In this chapter, the optimal way to select the number of knots for age-period-cohort analyses is investigated. A simulation study is used in order to do this. Two information criteria are compared to assess the degrees of freedom (synonymous with the number of knots) that are selected; the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

3.2. Introduction

When using splines in any framework, consideration must be given to the number and placement of the knots. It has been shown in a number of settings that model fitting is insensitive to knot selection under certain conditions [Harrell et al., 1988; Heinzl et al., 1996; Lambert et al., 2010]. In an age-period-cohort setting, a number of suggestions have been made for how many knots to use for each of the three functions. Heuer [Heuer, 1997] proposes that knots should be placed every 5 years for any of the given variables when using the yearly split data described in Section 2.4.1. Carstensen [Carstensen, 2007] suggests, in contrast, that an equal number of knots can be used for age, period, and cohort. This seems non-intuitive as the range of the period variable is often shorter than the range of age and cohort. The default degrees of freedom for the software available in R [Carstensen et al., 2008] written by Carstensen, is 5 degrees of freedom for each of the variables (leading to 4 internal knots).

This chapter is an attempt at verifying the optimal way to select the number of knots, and also to a degree, where these knots should be placed. A simulation study is used in order to do this. The simulation study compares a "true" underlying shape to that fitted by various models with differing degrees of freedom. Information criteria, namely the AIC and BIC, to compare the best model fit are then used to assess whether the "best"-fitting models provide

a good fit to the true underlying shape. In order to simulate realistic datasets, Finnish cancer registry data is used to inform the data generation process.

3.3. Simulation

An APC model using restricted cubic splines involves, either directly or indirectly, a decision on the placement for the knots. This decision is interwoven with the decision on how many knots should be used. This is simply because the fit to the data depends upon both the number of knots, and where the knots are placed. Two different fitted lines can be produced having used the same number of knots if the knots are positioned differently on each occasion. It is also the case that almost the same fit could be produced via two different spline models that have a differing number of knots if the placement of the knots is carefully chosen. This indicates that these two issues must either be considered simultaneously, or one of the factors must be kept constant whilst the other is investigated. The latter of these two approaches was deemed to be the most appropriate. The aim of the simulation is to assess the most appropriate way to select the number, and the placement, of the knots.

3.3.1. Simulated Datasets

In order to assess the performance of the methods for selecting the number, and the placement, of the knots it is essential to have a dataset in which the underlying "truth" is known. This "truth" provides a value to which the fitted values can be compared. This can only really be achieved by simulating a dataset of the appropriate form. That is, the dataset needed to have a known overall incidence rate of the disease whilst also having a known shape for each of the effects attributed to the three variables; Age, Period and Cohort. The level of complexity involved in creating such a dataset is increased due to the issue of identifiability that is a perennial issue when dealing with Age-Period-Cohort analyses. Therefore, it was decided to simulate data in order to simply fit an age-period model using the restricted cubic splines. The conclusions drawn surrounding the selection criteria and placement of the knots should be transferrable to the full APC model setting.

To give the simulated datasets a degree of realism, the datasets were based on fits to Finnish cancer data. In order to capture a variety of shapes and levels of incidence, three different types of cancer were used for this process. The three datasets that were used as the basis for the

simulated datasets were all from Finland, and were relating to cases of lung cancer, pancreatic cancer, and Hodgkin's lymphoma. Fractional polynomials were used as part of the process to create flexible shapes for the simulated functions.

Fractional polynomials for regression analyses were introduced by Royston and Altman [1994]. Fractional polynomials are an extension of regression modelling allowing a wide-range of functions to be selected for a continuous covariate. The fractional polynomial functions were forced to have a complex shape by selecting a fractional polymial forced to have 8 degrees of freedom. The 8 degrees of freedom relate to 4 selected parameters, and the 4 selected powers for those parameters. Therefore, a fractional polynomial function for x with 8 degrees of freedom is referred to as an FP(4) function and takes the form:

$$FP(4) = \begin{cases} \beta_0 + \sum_{i=1}^{4} \beta_i x^{p_i} & \text{if } p_i \neq p_j \ \forall \ i, j \\ \beta_0 + \sum_{i=1}^{k} \beta_i x^{p_i} + \sum_{j=1}^{m} \beta_j x^{p_r} * \ln(x)^{j-1} & \text{if } p_r \text{ is a repeated power}; k + m = 4, \end{cases} \tag{3.1}$$

where p_i are in the subset of powers $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. A power of 0 represents a logarithmic function for x. This allows a wide-range of shapes for the continuous representation of the variable. The process for selecting the powers and complexity has been laid out in the corresponding literature [Royston and Altman, 1994; Sauerbrei et al., 2006]. The process is similar to forward selection, whereby the simple linear model is fitted in the first instance and compared to the best-fitting FP(1) model. This process then continues until reaching a pre-defined limit of the complexity of the fractional polynomial function, or until the increase in complexity is not statistically significant at a pre-defined level (α).

The "true" fitted function was generated from the real Finnish datasets as follows:

(1) Collapsed Finnish cancer registry data were used for the number of cases, and relevant population risktime separated by age and period.

(2) A "forced" flexible fractional polynomial function was generated for age and period (forced to have 8 df). The powers $0, 1, 2, 3$ (where 0 is the log transformation) were used from the subset S for each of the terms so that a flexible shape could be produced.

(3) The period term was centred on a relevant reference period.

(4) A Poisson model with appropriate offset was fitted using the fractional polynomial terms.

(5) The shape for age and period used in this model were used as the "true" functions.

(6) From this model, the fitted rate was predicted. This was used as the "true" fitted rate, generated by the "true" shape of Age and Period.

3.3.2. Knot Placement

Two different approaches for the placement of the knots are considered. The analysed datasets are collapsed over age and period. The first approach places the knots equally across the range of the variable. That is, this approach ignores the fact that the distribution of the number of cases can vary across the values of the variables. For example, there may be an increased number of cases at the older age ranges for certain diseases leading to an increased amount of information (cases) for the higher ages. Therefore, it was decided to contrast the equally-spaced knot placement with a knot placement that accounts for the distribution of the number of cases. This method for the knot placement will be termed as the weighted knot placement, as the placement of the knots is weighted by the number of cases between the knots. The weighted knot placement places the knots so that there is an equal amount of cases between each of the knots. The two methods are compared separately in Figures 3.1 and 3.2; and are contrasted in Figure 3.3. It is clear from these figures that the placement of the knots can modify the shape of the underlying fit. This is particularly apparent for the younger ages, where there are a lack of knots for the weighted knot placement approach.

Figure 3.1 shows the placement of the knots for the Age variable in the Finnish colon cancer case if four internal knots are used (six knots overall; five degrees of freedom) and a method of equally spacing the knots is employed. This uniform knot placement leads to the four internal knots being placed at equal intervals across the range of the Age term. The histogram along the foot of the graph shows the distribution of the number of cases over the age range. As would be expected, the majority of the cases are confined to the larger age values. The uniform knot placement fails to take into account the amount of information (the number of cases) that are between each of the pairs of knots.

Figure 3.2 represents the knot placement when the knots are placed according to the distribution of the cases. The centiles of the age value for the knot placements are calculated according to a frequency weighting for the number of cases. This leads to the knots being accumulated around the larger age values because this is where the majority of the information

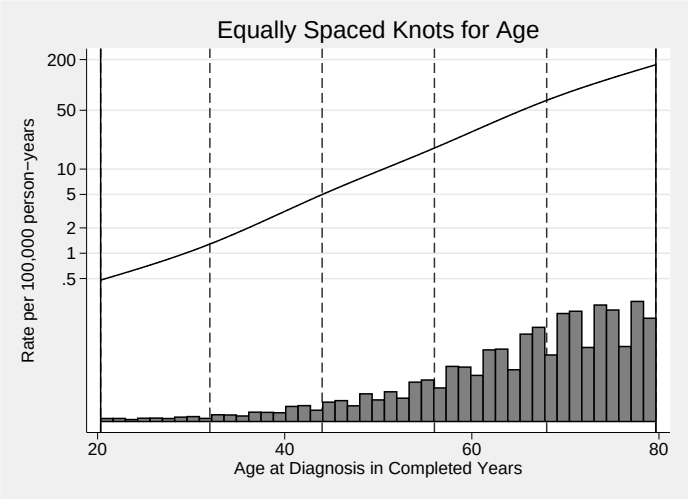

FIGURE 3.1. Fitted function for the effect of age with equal knot placements.

is contained. It can be seen from Figure 3.2 that the first internal knot is placed at around the age value of 57, whereas in Figure 3.1 the third internal knot is placed at roughly this point. This inevitably leads to a difference in the fitted values that are obtained for each of these knot placements. In the weighted knot placement, it is the number of cases that are equally distributed between each of the knots. It has been suggested that this knot placement is of benefit in that the knots are placed so to capture any deviations when a lot of information is present, and place less weight on deviations when there is less information to corroborate the deviation from the trend.

The two figures are combined in Figure 3.3, which gives the comparison of the two knot placements for the age effect. This figure highlights the difference in fit between the two functions where there is less information, that is, at the lower end of the age range. Having no knots at all before age 57 forces the cubic function to be less flexible in the younger ages for the weighted knot placement. However, in this example the added benefit of having an increased number of knots for the older age range is not directly apparent from the graph, as the two lines appear to overlay almost perfectly from age 50 onwards. If fewer overall knots were selected or there was a greater deviation from linearity later in the shape of the age curve, this may have led to a greater difference being observed.

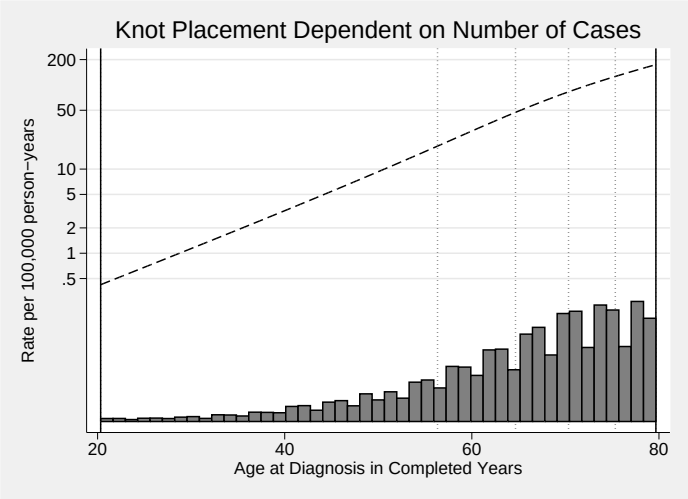

FIGURE 3.2. Fitted function for the effect of age with weighted knot placements.

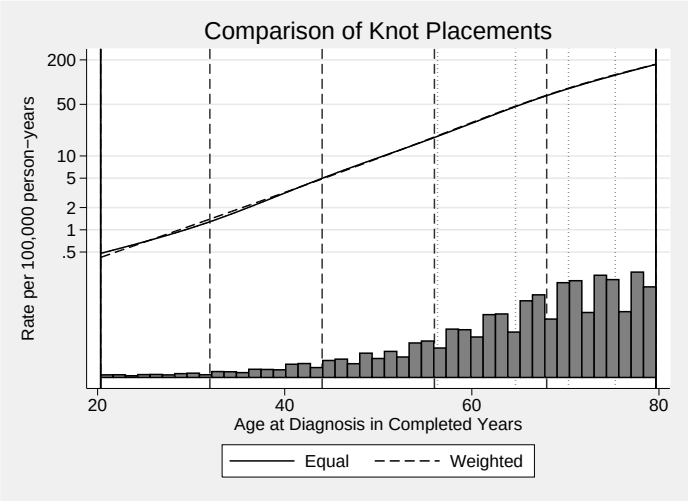

FIGURE 3.3. Comparison of the equal and weighted knot placements.

3.3.3. Number of Knots

A decision about the number of knots to use for a given variable is a decision that is closely related to the placement of the knots. In the case of fitting an APC model using restricted cubic splines, a weighted knot placement could provide a similar fit to an equally-spaced knot placement by using fewer knots. Generally, the data that APC models are fitted to have a large amount of information and, therefore, the expense of a few additional unnecessary parameters in the model is minimised. However, this may not always be the case and it is good practice

to use as few parameters as possible to efficiently summarise the data. It is possible that a selection criteria, such as the AIC [Akaike, 1973] or BIC [Gelfand and Dey, 1994], could be used in order to determine the appropriate amount of knots to be used in a given example.

3.4. Methods

3.4.1. Varying the Size of the Dataset

As part of the investigation into the placement of the knots in Age-Period-Cohort analyses, it was decided that various population sizes should be considered in order to cover a full range of real-life scenarios. These models may well be fitted to much larger populations than Finland, and also may well be used for regional analysis - which could result in a smaller population size than Finland being considered. Rather than doing a detailed analysis on a range of populations, it was possible to use the datasets created for this simulation, by multiplying the number of cases, and population risk-time, by a given factor to mirror real-life populations with the same underlying rate of disease. The factors selected were 0.1 to reduce the size of the population, and 10 to increase the size of the population.

3.4.2. Simulation Process

(1) In each of 500 simulations, the number of cases were drawn from a Poisson distribution with a mean of the "true" number of cases for each age-period combination.

(2) Using this as the basis for the dataset, a spline generating package (`rcsgen` in Stata) were then used to generate spline functions with varying degrees of freedom for the age and period terms. Models were then fitted to the data with each combination of degrees of freedom from 3 to 15 for both Age, and Period. This led to 169 models being fitted for each of the three dataset sizes for each run of the simulation.

(3) A weighted placement for the knots was also used by using `rcsgen` with a frequency weight (`fw`) equal to the number of cases for that particular simulation. This led to a further 169 models for each of the dataset sizes for every run of the simulation.

(4) The fitted functions for Age, and Period as well as the AIC and BIC values were stored for each combination of the Age and Period degrees of freedom.

3.4.3. AIC and BIC

Information criteria can be used to assess model fit and can be used to select the best-fitting model from a compared set of models. The information criteria attempt to strike a balance between model fit and simplicity; often referred to as attempting to select a parsimonious model. A recent appraisal of the two most common information criteria used; Akaike's An Information Criteria (AIC) [Akaike, 1973] and the Bayesian Information Criteria (BIC) [Gelfand and Dey, 1994], is given by Burnham and Anderson [2004]. There is literature examining which of these approaches should be used in a given situation and criticising the BIC for favouring models that are too simple [Weakliem, 1999]. However, both criteria are applied to analyse which of the approaches performs well in the simulation; rather than concentrating on the philosophical and mathematical reasoning for preferring either of the criteria.

The formula for the AIC is given by:

$$AIC = -2\ln(L) + 2k, \tag{3.2}$$

where L is the maximised value of the likelihood, and k is the number of parameters in the model. The corresponding formula for the BIC is given by:

$$BIC = -2\ln(L) + \ln(n) * k, \tag{3.3}$$

where n is the number of observation points; the total number of cases was used as the value of n for the analyses in this chapter. The terms including the number of parameters act as a penalty term, to punish models with excess parameters which do not sufficiently decrease the log-likelihood. The difference between the two approaches is the size of this penalty term. On the basis that $\ln(n)$ is greater than 2 where $n \geqslant 8$ (that is, when there are at least 8 observation points), the BIC will be more likely to select a simpler model in all practical applications.

In practice, both criteria have been used simultaneously when comparing spline techniques of various forms [Molinari et al., 2004]. They have also been used in unison in stepwise techniques when using splines in non-parametric regression [Doksum and Koo, 2000]. There are other examples when using splines when one of the two approaches has been favoured [Yoshimoto et al., 2003]. Furthermore, simulations have been carried out to investigate various spline, and smoothing, techniques where the AIC has been applied to select the most appropriate

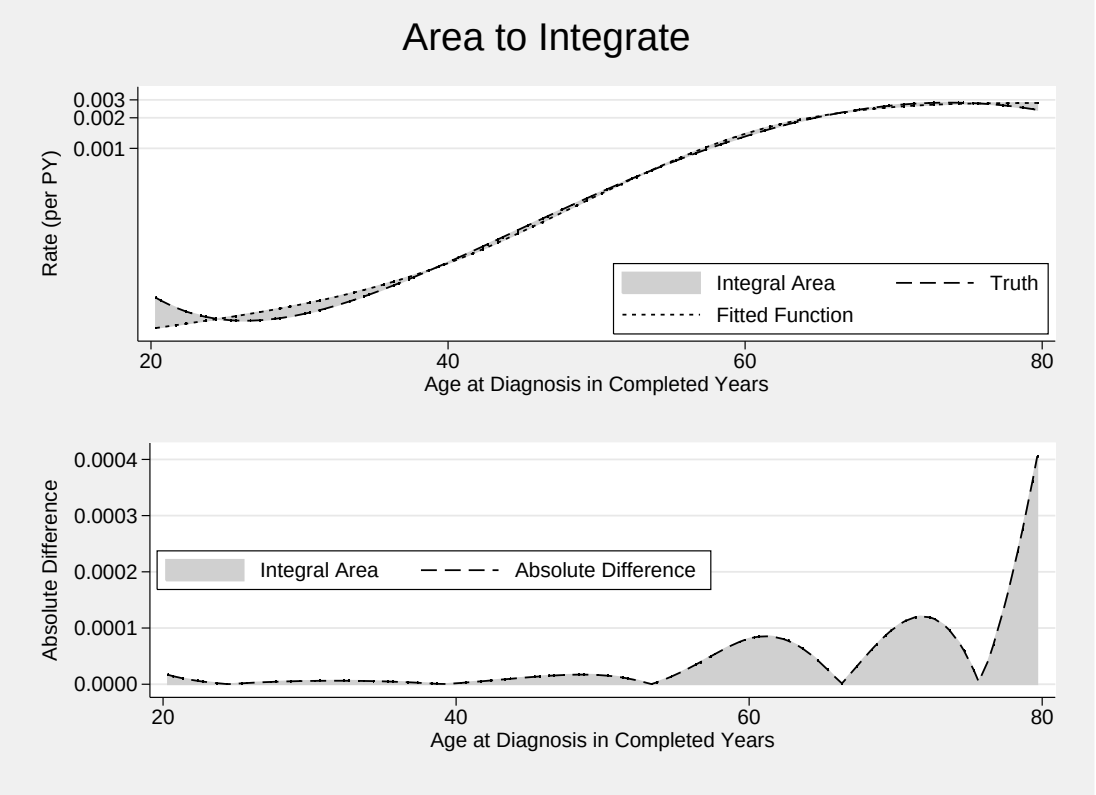

FIGURE 3.4. Illustration of the difference in area to be calculated by numerical integration.

models [Govindarajulu et al., 2009]. There are also examples of applications to real data to compare smoothing techiques for Cox models using the AIC [Govindarajulu et al., 2007].

3.4.4. Assessing the Fit of the Functions

The main method of comparison to compare the fit of the functions is illustrated in Figure 3.4. The absolute area difference between the "true" and fitted functions is calculated using numerical integration techniques. For the age functions, this integration is performed on the rate scale, as the age function is given the rate dimension by including the constant term as recommended by Carstensen [2007], and as performed in the analyses in the previous chapter. In contrast, the integration for the period function is performed on the rate ratio scale. This alters the interpretation of the area differences that are calculated in each case.

Clearly, the simplest approach is to assess which of the fitted functions gives the smallest area difference when comparing to the "truth". However, it is more informative to plot the area difference across the range of the degrees of freedom. This provides an assessment of how

much a worse fit (compared to the best-fitting model compared) is achieved by fitting greater or fewer degrees of freedom for the age and period terms.

3.4.5. Other Methods of Comparison

As an extension to the main method of comparison between the different models, three extra comparison methods were considered. Firstly, the number of turning points for any fitted curve can be calculated by looking at changes in sign for the derivative of the spline functions. The "true" number of turning points can also be obtained from calculating the derivates of the fractional polynomials. If a fitted spline function captures the true shape of the fractional polynomial, it should at the very least have the same number of turning points (as well as these turning points being placed in the same place). Histograms of the number of turning points for the 500 simulations for different combinations of degrees of freedom for age and period were also used as a method of comparison. The spline functions selected by the information criteria can also be compared in terms of the number of turning points in the fitted functions they select.

The other further methods of comparison also concentrate on the difference betweeen the two information criteria in terms of the degrees of freedom selected. Firstly, the value that minimises the difference in area can be considered the optimal degrees of freedom for that comparison. It is then possible to compare the values for the degrees of freedom selected by the information criteria to the "optimal" degrees of freedom. Again, these comparisons are done by using histograms to show the distribution for each of the criteria. Finally, it is also of interest to compare the difference between the degrees of freedom selected by the AIC compared to that selected by the BIC. Due to the harsher penalty term that is employed when calculating the BIC value (see Equation (3.3)), the BIC will tend to select a simpler model than the AIC. Histograms of the difference in degrees of freedom for the various simulation scenarios are given.

3.5. Results

Firstly, the results are given for a single simulation (Section 3.5.1) for one of the scenarios in order to clarify the process of the simulation and the results that are obtained in Section 3.5.2. The scenarios are then run for 500 simulations each (Section 3.5.2) and the results are given graphically to assess the performance of the AIC and BIC, and to compare the two methods

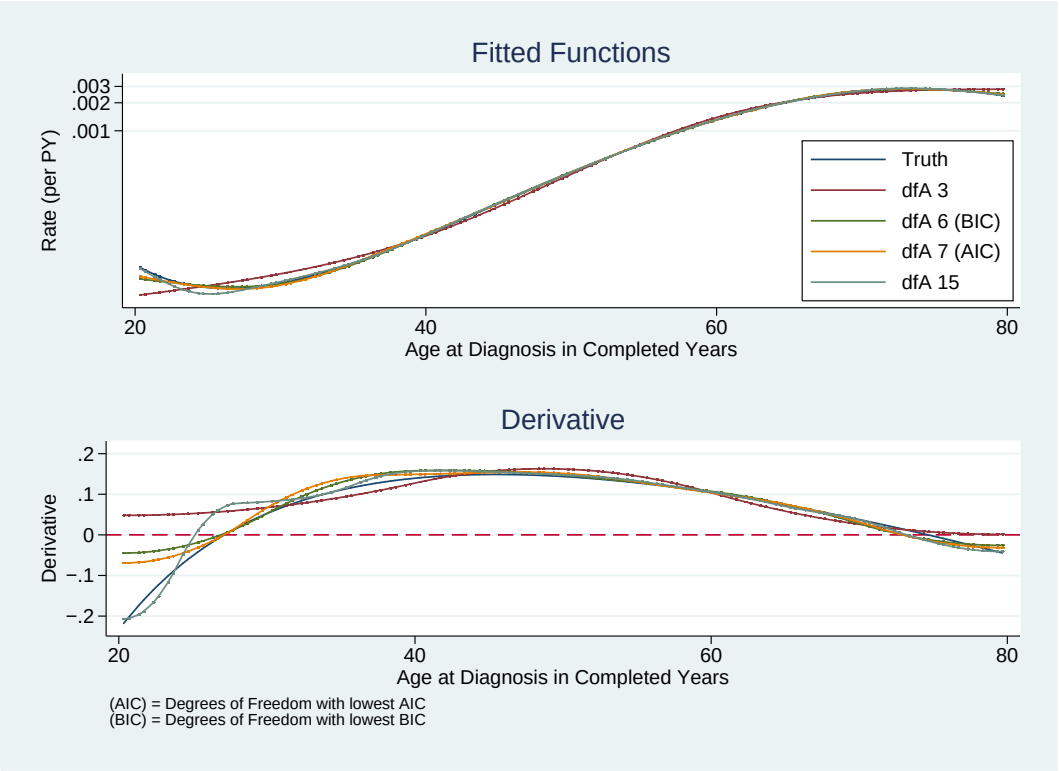

FIGURE 3.5. Results of the single simulation for the age curve for lung cancer (equal knot placement).

for knot placement. The "true" shape generated from the fractional polynomials for each of the cancer sites are given in Figures 3.9, 3.15 and 3.20 in Section 3.5.2.

3.5.1. Single Simulation

A single run of the simulation was performed for lung cancer. Taking a more detailed look at a single run of the simulation can lead to further insight into the reasons behind the observed differences between the fitted values and the "true" simulated shape. The single run can be used to highlight examples of underfitting and overfitting when varying the number of knots used for the splines. However, this is only a single run of the simulation and may not be representative of the overall fit. The full simulation results given in Section 3.5.2 are used to assess this for each of the cancer sites.

3.5.1.1. *Equal Centile Knot Placement*

Figure 3.5 shows the resulting output from a single simulation for the age effect. The simulated data were based on the underlying shape for lung cancer patients in Finland. However, the

upturn in the "true" age effect for the youngest ages is in fact an artefact from using fractional polynomials to try to capture the underlying shape. Using this as the "truth" leads to a more complicated shape for the effect of age and, therefore, provides a better test for the flexibility of the spline functions. The "true" fractional polynomial function for age is given in the graph alongside the fitted functions for a selection of degrees of freedom for the splines for Age (df=4 was chosen for Period for each of these curves; the AIC value selected in this example). It is clear that using 3 as the degrees of freedom leads to an underfitting of the data. The line shows a lack of fit for the youngest and oldest ages. At the other end of the spectrum, using 15 degrees of freedom seems to suggest a case of overfitting for this simulated dataset; this is particularly highlighted in the graph for the derivative of the fitted function. It seems that a local minima is accentuated for age 25 in this simulation and that using too many knots has led to this being captured by the spline function. However, it should be noted that the differences are small in each case. The other lines on the graph indicate the fit chosen by the two information criteria. The AIC selects a more complicated shape, which seems to fit better when appraising the curves by eye. The area between the curves will be used to assess these differences analytically when the simulation is carried out in full.

Figure 3.6 shows the accompanying graph for the period effect. Again, it should be noted that the differences between the curves are quite small. The reference period of 1980 is indicated by the hollow circle. The "truth" from the fractional polynomial for the period effect is compared to the fitted functions according to the knot selection for the AIC and BIC as well as a knot selection that clearly overfits the data. Using 15 degrees of freedom for the period term clearly leads to local minima and maxima being "picked up" by the fitted function. These are features of the simulation process and it is unlikely that it would be desirable for a selection criterion to prefer a function that captured these local effects. It can be seen that the shape is fairly simple for the period effect in this case and this is reflected in the values for the degrees of freedom (df) selected by the AIC and BIC (4 and 3 respectively).

3.5.1.2. *Weighted Knot Placement*

The same single simulation was carried out using a weighted knot placement for the positioning of the knots. The weighted knot placement places the knots so that an equal number of cases

FIGURE 3.6. Results of the single simulation for the period curve for lung cancer (equal knot placement).

are contained between any given pair of knots. This is in contrast to equally spacing the knots at appropriate centiles of the data.

Figure 3.7 highlights a major issue if deciding to use a weighted knot placement as opposed to knots placed at equal centiles. Each of the selected fitted functions fails to capture the true shape of the age curve for the youngest ages. This is purely on the basis that there are fewer cases at the younger ages, and therefore, the knots are concentrated around the oldest age groups (as was the case in Figure 3.2). For example, if an equal knot placement had been used in the case of 5 degrees of freedom for the Age curve, the knots would be placed at (20.33, 32, 44, 56, 68, 79.67). However, if a weighted knot placement was used instead the knots are placed at (20.33, 57.33, 63.67, 68.33, 73.33, 79.67). Therefore, using the weighted knot placement, it is not possible to capture a complex shape in areas where there is less "information". This property can be considered to be both desirable and undesirable for various reasons. It is desirable in that the shape of the curve is allowed greater complexity where more information

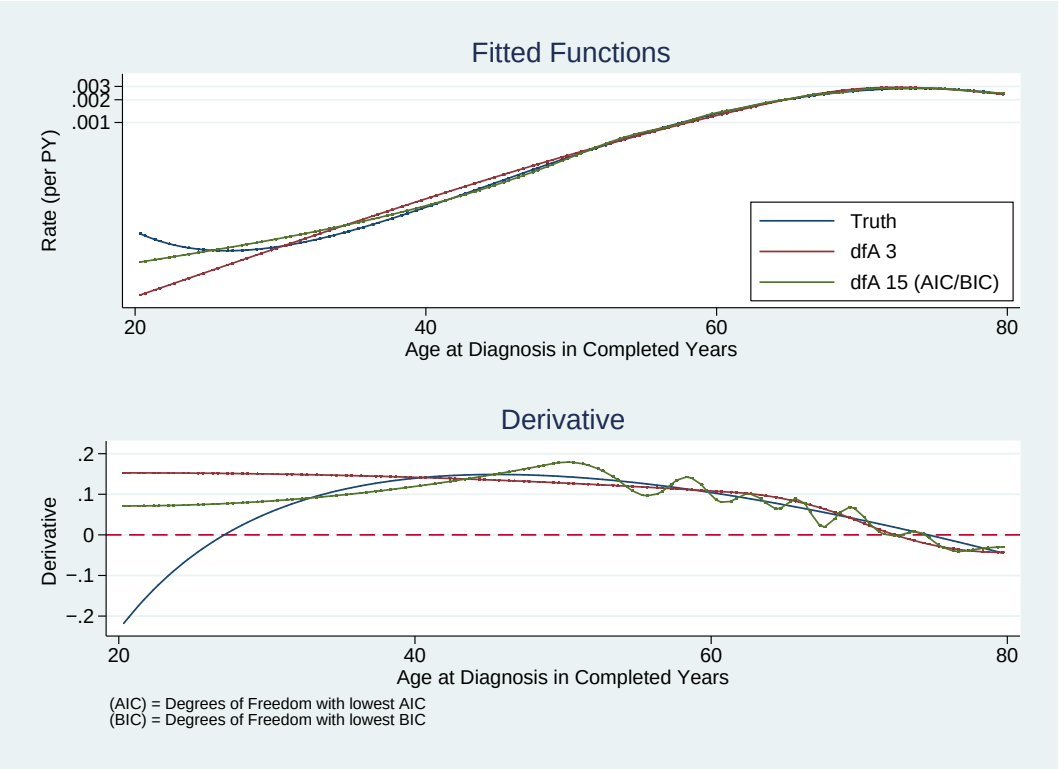

FIGURE 3.7. Results of the single simulation for the age curve for lung cancer (weighted knot placement).

is available and a simpler shape is enforced if there is less data to dictate the shape. However, if the data indicates that there is a true shape even when the data is sparse, it may be of interest to still capture this shape with the spline function. In the example given in Figure 3.7, it is clear that the true shape cannot be captured for the youngest ages even when increasing the degrees of freedom to 15. The first internal knot for the age effect with df=15 is still as high as 48.67, indicating the weight of information that is contained for the older ages. However, this shape was captured in Figure 3.5 when using the equally-spaced knot placement.

In this example, the knot placement for the period term does not vary too much if a weighted knot placement is used rather than the equally-spaced knot placement. Given the distribution of cases over the period 1953-2007, this is hardly surprising. On that basis, a similar fit can be achieved irrespective of the knot placement. The final consideration is the fit of the weighted knot placement if the cohort term had been fitted as well in this analysis. On the basis that the later cohorts are made up of the youngest age-groups, this will lead to fewer knots being placed

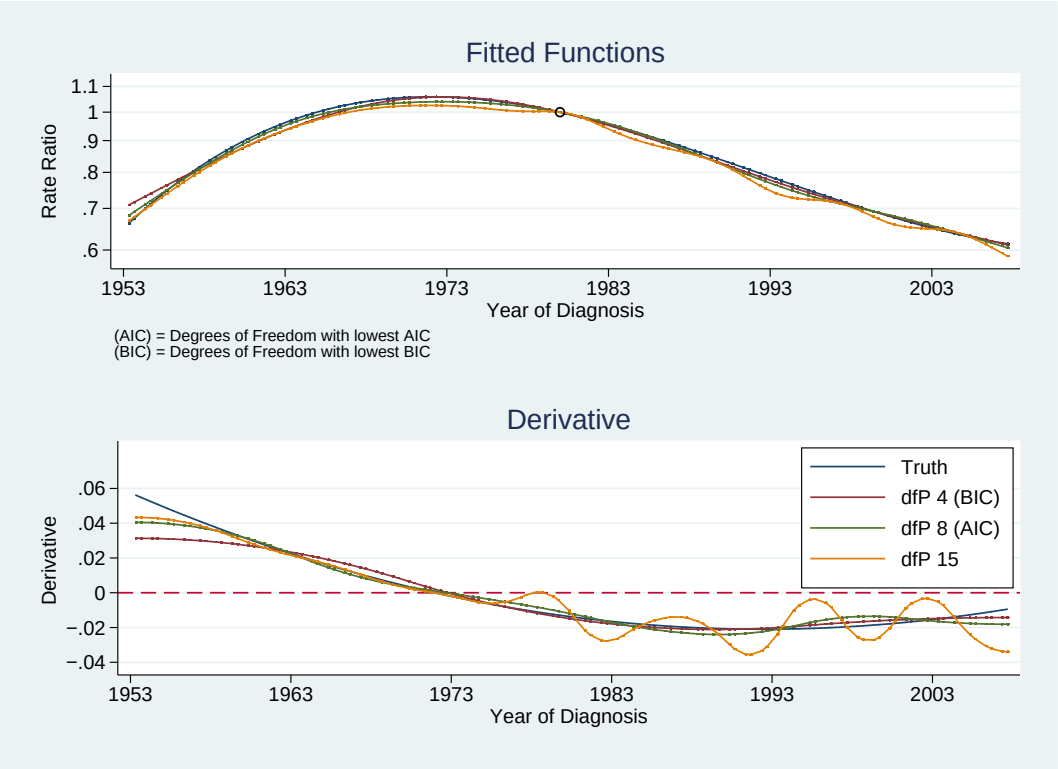

FIGURE 3.8. Results of the single simulation for the period curve for lung cancer (weighted knot placement).

towards the end of the cohort function if the weighted knot placement is used (for the majority of cancer sites where incidence increases with age, and a diagnosis is fairly rare at younger ages). This may have important implications when considering projections of the age-period-cohort models (see Chapters 4 and 5), and it may be another negative of the weighted knot placement approach.

3.5.2. Full Simulation

The results of the full simulation are reported for each of the three cancer sites separately. The other methods of comparison (Section 3.4.5) are exemplified for at least one of the cancer sites to give further insight than the main method of comparison.

3.5.2.1. *Simulations based on Lung Cancer*

Figure 3.9 gives the fractional polynomial shapes for age and period that were selected using the lung cancer dataset. The simulated shape for the age effect has two turning points, and one of these turning points is in an age range where it is likely a small number of cases will

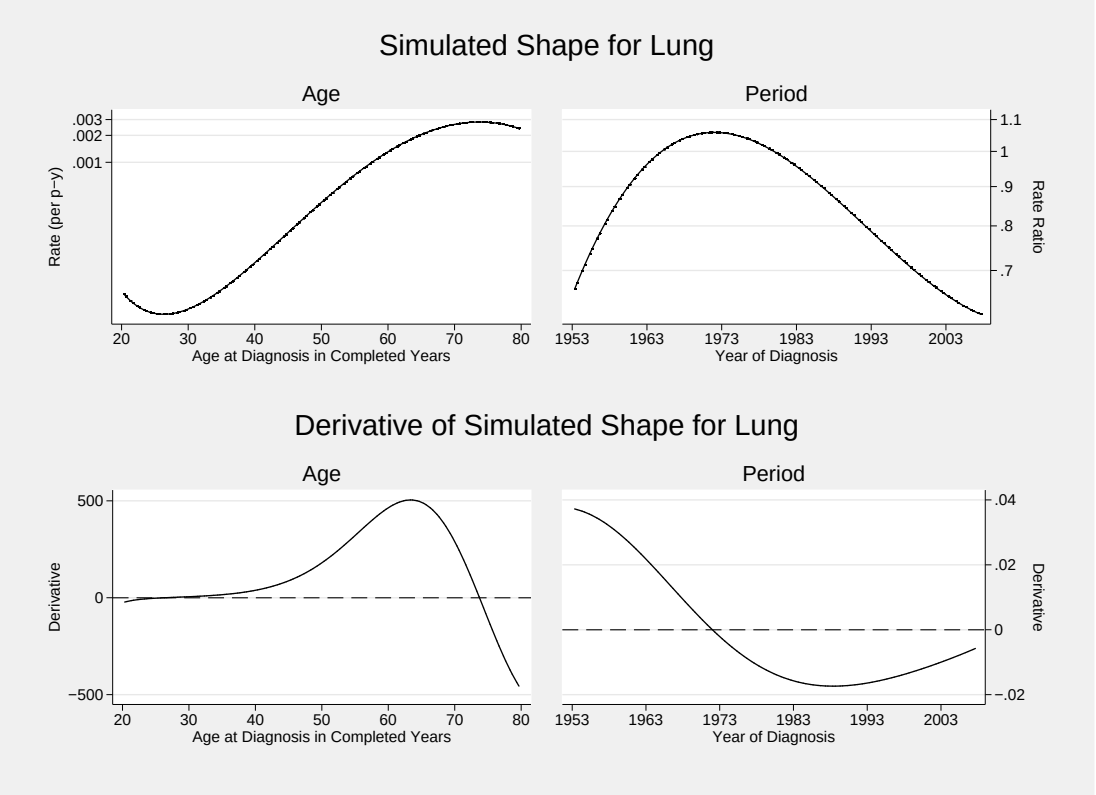

FIGURE 3.9. The true shape generated using fractional polynomials for lung cancer.

be simulated. The shape for period over calendar time has a single turning point, and is a relatively simple underlying shape for the cubic splines to capture. In the lower half of Figure 3.9 the derivatives of the fractional polynomial functions are plotted to indicate the turning points and smoothness of the functions. The age curves are given on the rate scale and are given per person-year. The period curves are given on the rate ratio scale.

Figure 3.10 shows the results of the 500 simulations for the dataset based on lung cancer. The left-hand graphs correspond to those relevant to the Age curve, whereas the right-hand graphs are related to the Period curves. The top-row relate to the simulations performed with an equally-spaced knot placement, and the bottom-row of graphs are relevant to the comparisons for the weighted knot placements. The dashed vertical line indicates the average (mean) value of the degrees of freedom selected by the BIC, and the solid vertical line indicates the average degrees of freedom selected by the AIC. The df selected by the AIC is consistently higher than that selected by the BIC. The non-weighted curves for Period show the pronounced shape that

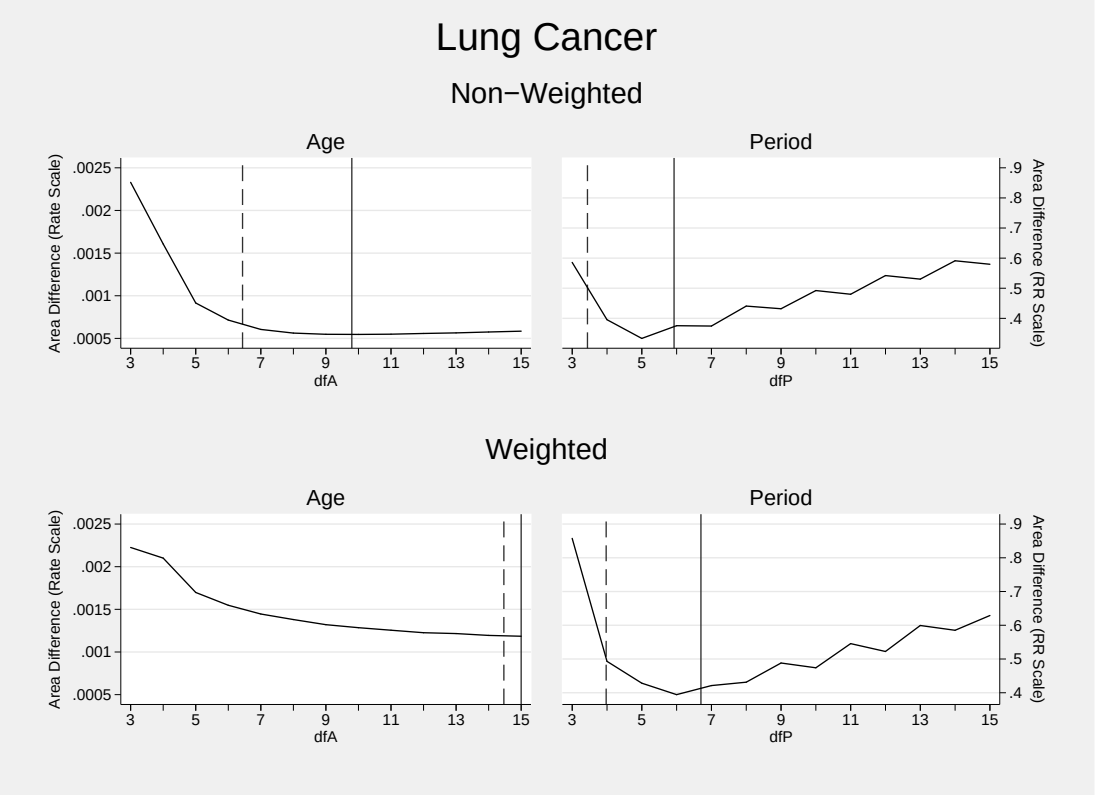

FIGURE 3.10. Results of the full simulation for lung cancer (Factor=1).

is to be expected; the area difference is higher when the shape is under-fitted, and it is also higher when the shape is over-fitted. The minimum of the curve falls at around 5 degrees of freedom. The period curve is fairly similar when a weighted knot placement is used for this example. This was seen in the single simulation in the previous section and can be explained by the distribution of the number of cases across the period. However, an interesting feature of the graph for period is the jagged nature of the curve. This "zig-zagged" shape was an unexpected outcome of the simulation results and required further investigation. In the following section (Section 3.5.3), this phenomenon is investigated. In the case of the age curves in Figure 3.10, there is a significant difference between the values selected for the AIC and BIC for the non-weighted and weighted knot placements. This is due to the fact that, when using the weighted knot placement, the shape of the curve for the younger ages cannot be fully captured due to the lack of knots. Therefore, the information criterion require a greater number of knots to try to improve the fit for the younger ages. The simulations were only carried out using between 3

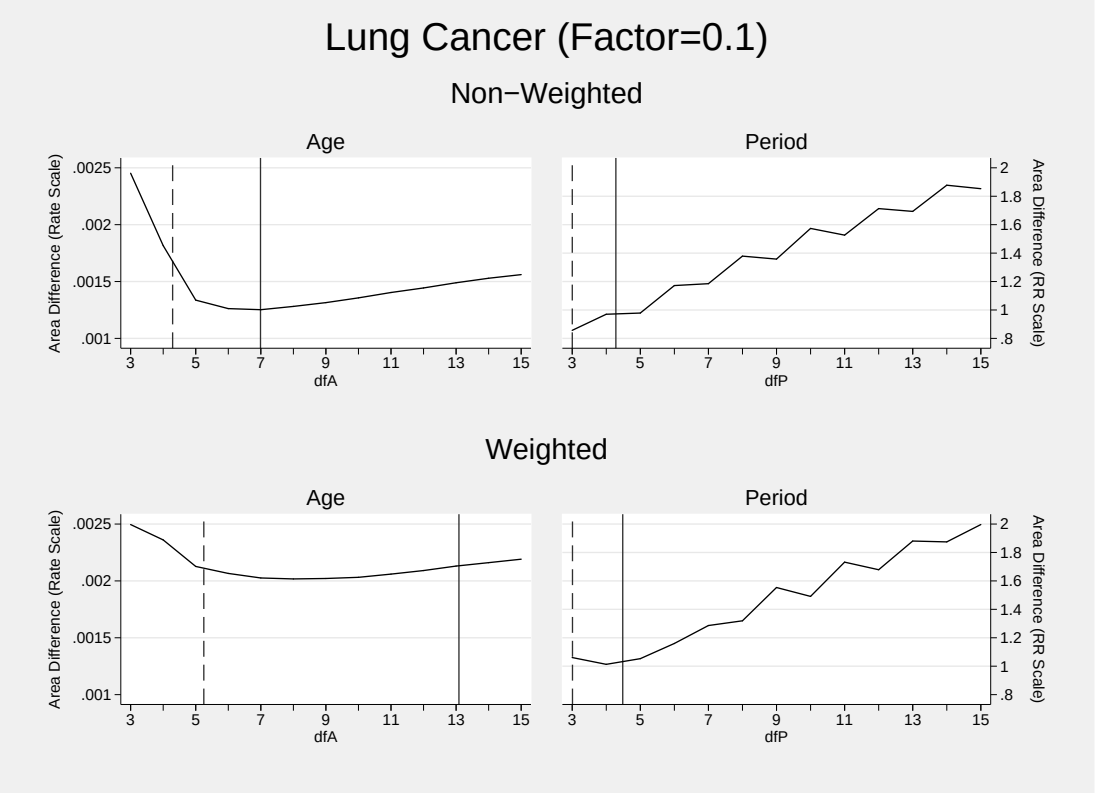

FIGURE 3.11. Results of the full simulation for lung cancer (Factor=0.1).

and 15 degrees of freedom and an even greater degrees of freedom than 15 would be required in order to capture this early curvature when using a weighted knot placement because a knot is required at a younger age.

Figure 3.11 contains the results of the simulation when the cases and population size have been scaled down by a factor of 10. With less information, and smaller values for the number of cases, a simpler spline function is selected on average by the two selection criteria. From the values on the scales of the graphs, it is also clear to see that the corresponding spline functions fit less well on average than in the case where more information was available. The difference between the weighted and non-weighted knot placements is consistent with the observations drawn from the case with no scaling factor on the population size.

Figure 3.12 contains the results of the simulation when the cases and population size have been scaled up by a factor of 10. This leads to a scenario with a greater amount of information, and larger values for the number of cases. The corresponding values for the average AIC and

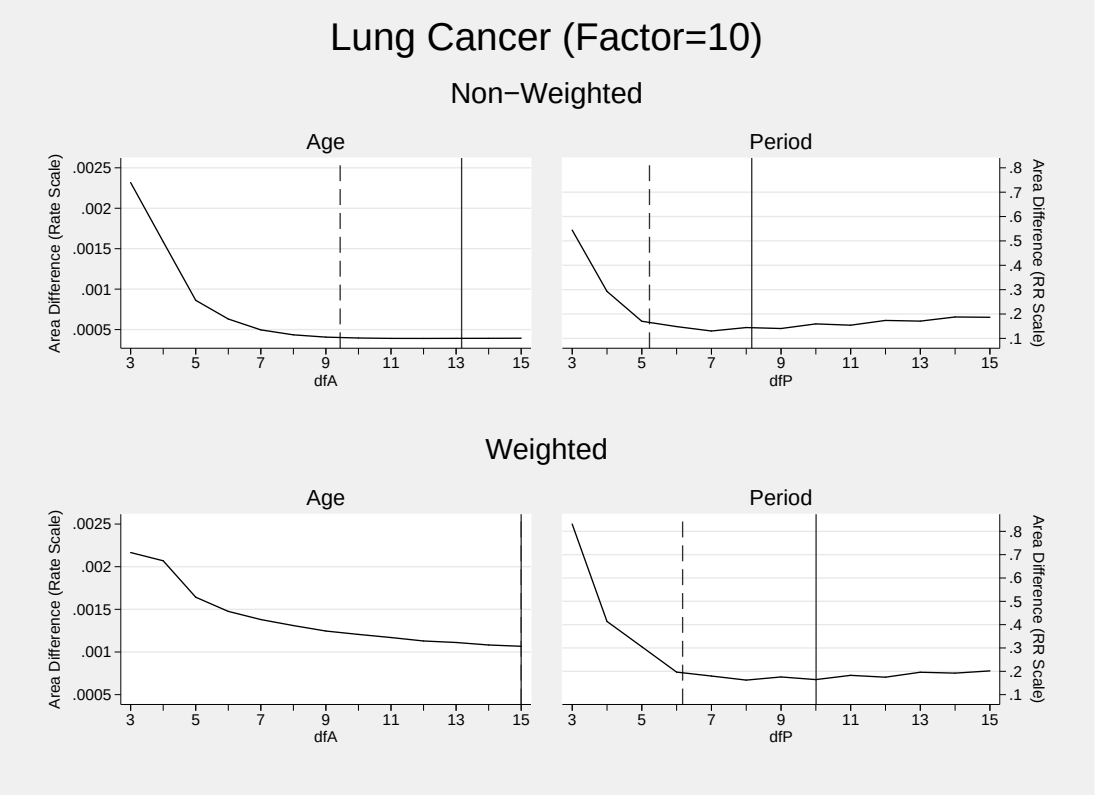

FIGURE 3.12. Results of the full simulation for lung cancer (Factor=10).

BIC knot selections are consequently higher. With a greater amount of information to dictate the shape of the curves, it is less likely that there would be a sufficient clustering away from the true shape in the simulated data that would lead to undulations in the fitted curves. In this case, there is very little overfitting for the higher df selections, which is apparent from the curves being flatter after reaching their minimum. The issue of the weighted knot placement not being able to capture the shape of the age effect in the youngest ages is even more apparent when increasing the size of the data. For the age curve and weighted knot placement in Figure 3.12, the df selection for both of the selection criteria are on average 15 (the maximum considered).

Figure 3.13 contains the histograms of the number of turning points of the fitted function for the age curve for each of the 500 simulations under the different scenarios. A histogram is given to show the turning points for the curve using the two most extreme degrees of freedom values (3 and 15), as well as for the curves selected by the two information criteria. The true number of turning points for the Age curve is 2 (as illustrated in Figure 3.9). Using 3 degrees of

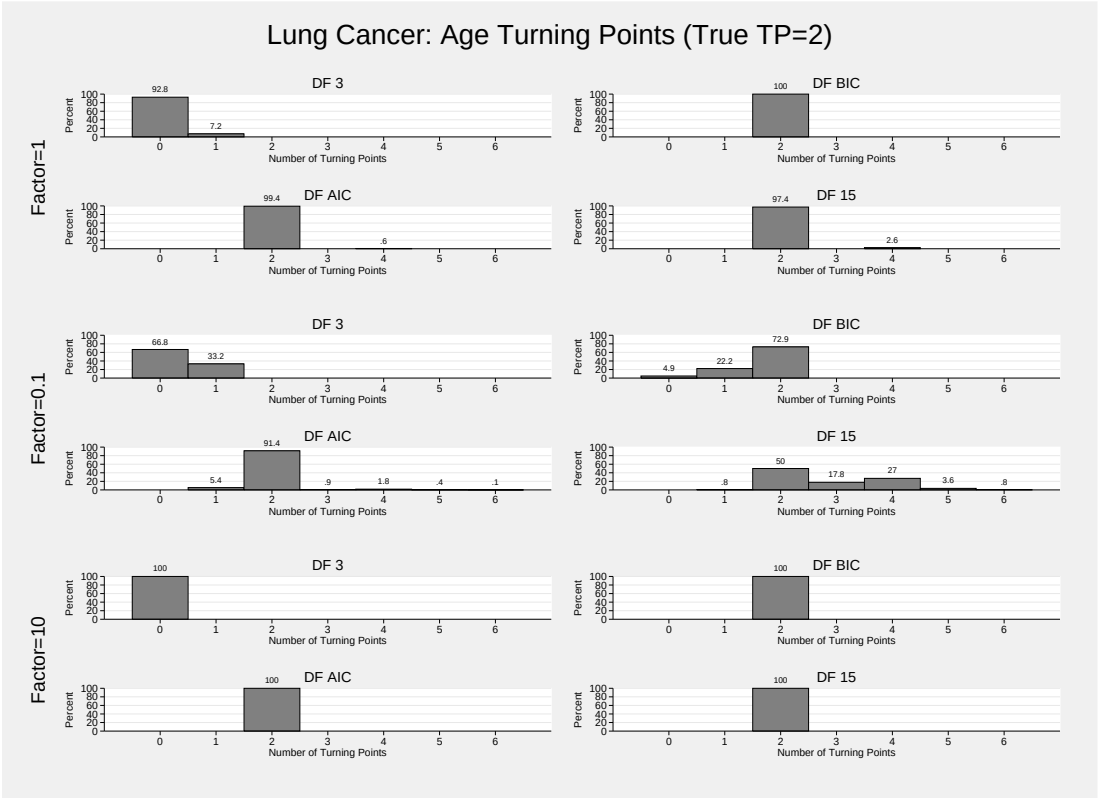

FIGURE 3.13. Histograms showing the number of turning points for the age curves. Lung cancer.

freedom for the age curve leads to the true number of turning points to be underestimated for all of the scaling factor scenarios. Using 15 degrees of freedom does not produce an overestimate for the number of turning points in the case when the Factor is set to 10. Also, in the case when Factor is set to 1, there is only a small proportion of the 500 simulations (2.6%) that produce an overestimate. However, for the case when a reduction of the size of the population is used (Factor=0.1), it is clear that the overfitting is leading to an increase in the number of turning points in almost half of the simulations (49.2%).

It is also clear from the plots in Figure 3.13 that using the degrees of freedom selected by the BIC never leads to an overestimate of the number of turning points in these scenarios. However, in the scenarios where there is less information (Factor=0.1, 1) the AIC degrees of freedom can produce a curve with extra turning points in a small proportion of the 500 simulations (3.2%, and 0.6% respectively). The two information criterion estimate the number of turning points correctly for the vast majority of cases when the scaling factor is equal to 1 (AIC 99.4%, BIC

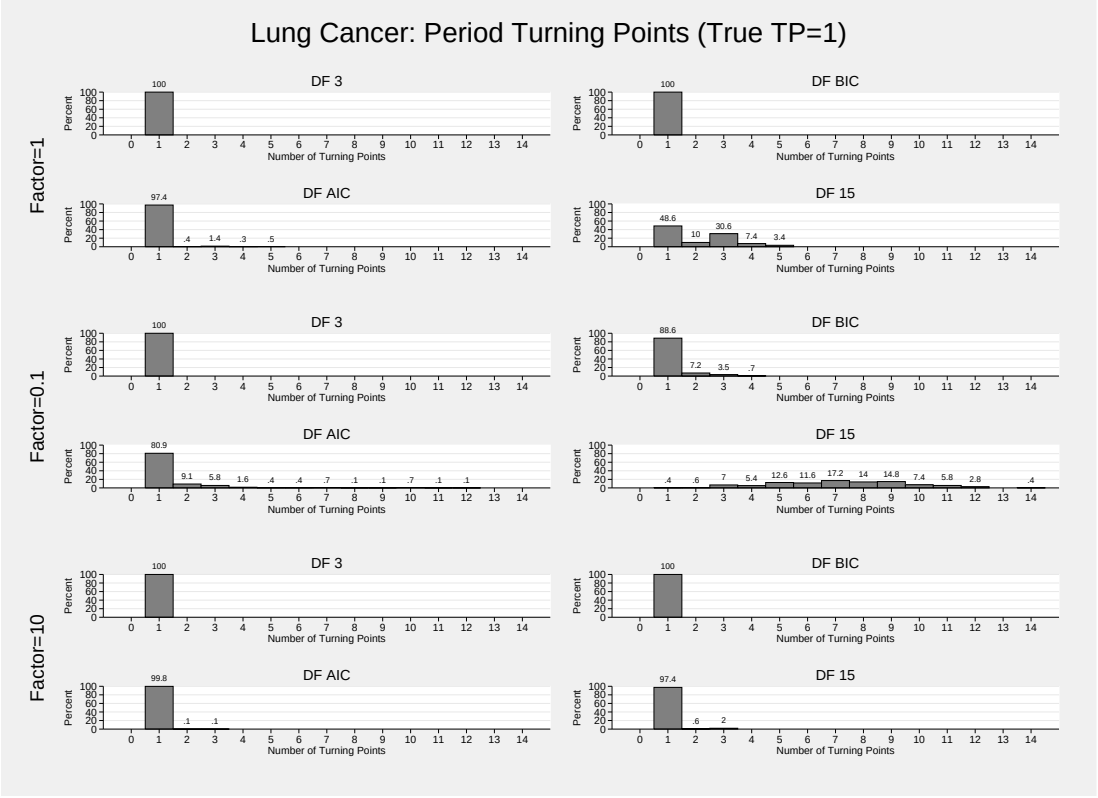

FIGURE 3.14. Histograms showing the number of turning points for the period curves. Lung cancer.

100%) and when the factor is equal to 10 (AIC 100%, BIC 100%). In the case where there is less information (Factor=0.1), the degrees of freedom selected by the BIC underestimates the true number of turning points in a higher proportion of the simulations (27.1%). This suggests that the BIC is selecting an overly simplistic function in these cases.

Figure 3.14 gives the corresponding scatter of histograms for the period curve for lung cancer. The simulated shape for period has a single turning point as illustrated in Figure 3.9. At least one turning point is picked up by all of the compared degrees of freedom values, even for the simplest model with 3 degrees of freedom. Unless there is a large amount of information to dictate the shape (Factor=10), the models using 15 degrees of freedom tend to overfit the model for some of the simulated examples according to the number of excess turning points for the fitted functions. On the whole, the information criteria estimate a similar proportion of the simulations to have the correct number of turning points. There is some evidence that the function selected by the AIC tends to have a severely overfitting shape in a very small

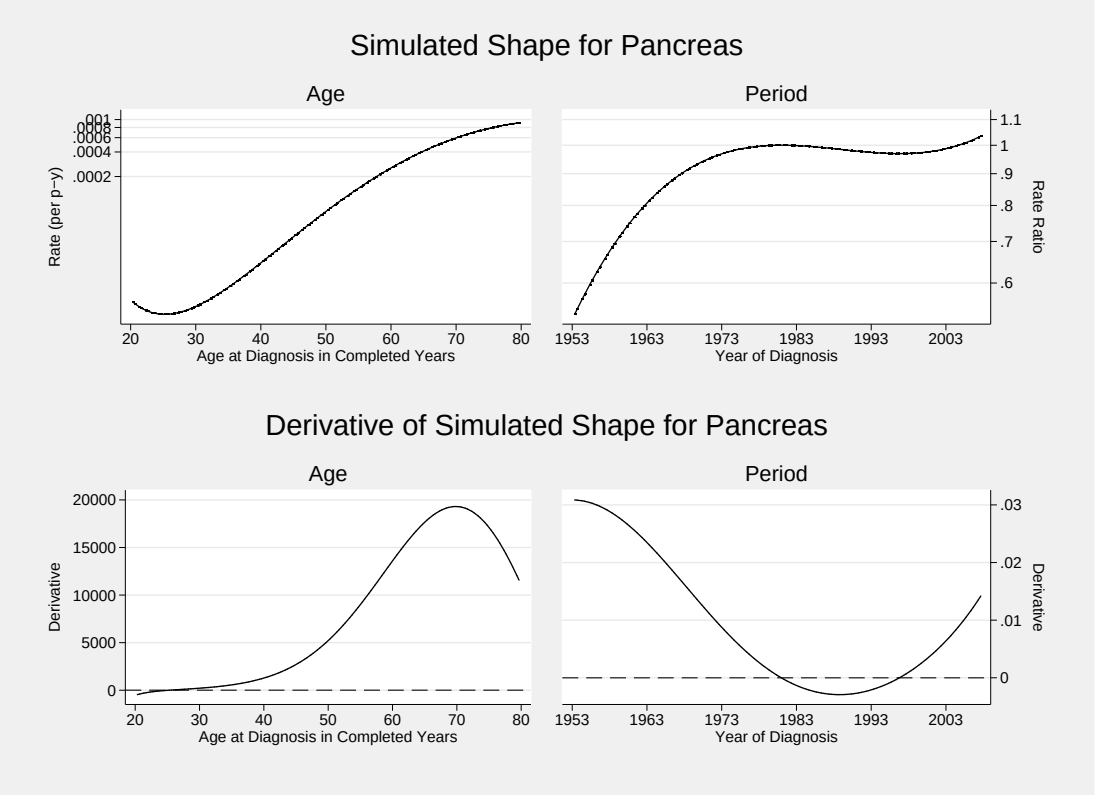

FIGURE 3.15. The true shape generated using fractional polynomials for pancreatic cancer.

proportion of cases, with up to 12 turning points captured when using the smallest dataset (Factor=0.1).

3.5.2.2. *Simulations based on Pancreatic Cancer*

Figure 3.15 gives the fractional polynomial shapes for age and period that were selected using the pancreatic cancer dataset. The simulated shape for the age effect has a single turning point at an age range where it is likely a small number of cases will be simulated. The shape for period over time has two fairly subtle turning points and therefore may require an increased number of degrees of freedom in order to capture the true shape.

Figure 3.16 shows the results of the 500 simulations for the dataset based on pancreatic cancer. The results are similar to those observed for lung cancer, and the expected shape for the area difference curves is seen. Using too few degrees of freedom leads to a greater area difference due to underfitting, whereas using too many degrees of freedom leads to local features being captured, which also increases the area difference to the true shapes for age and period. The

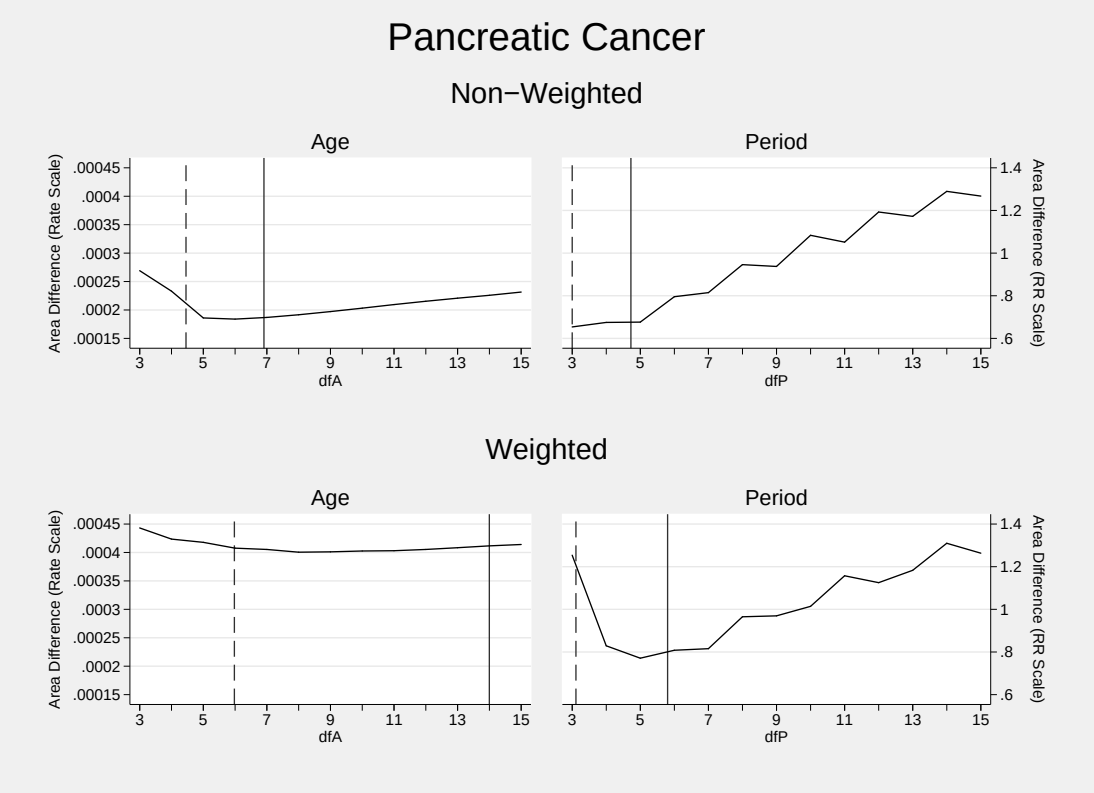

FIGURE 3.16. Results of the full simulation for pancreatic cancer (Factor=1).

weighted knot placement suffers due to the early turning point that is apparent for the Age curve (where there are fewer number of cases).

Figure 3.17 shows the results of the 500 simulations for pancreatic cancer when the cases and population size have been scaled down by a factor of 10. It can be seen that reducing the relative size of the population results in a simpler shape being preferred by the information criterion. Reducing the size of the population increases the opportunity for local features to be prominent, meaning that the higher degrees of freedom are more likely to overfit compared to the "true" shapes of Age and Period.

Figure 3.18 shows the results of the 500 simulations for pancreatic cancer when the cases and population size have been scaled up by a factor of 10. It can be seen that increasing the relative size of the population results in a higher degree of freedom being selected by the information criterion. The unexpected issue of the weighted knot placement for the Age curve is again apparent, with both information criteria selecting 15 as the degrees of freedom in each

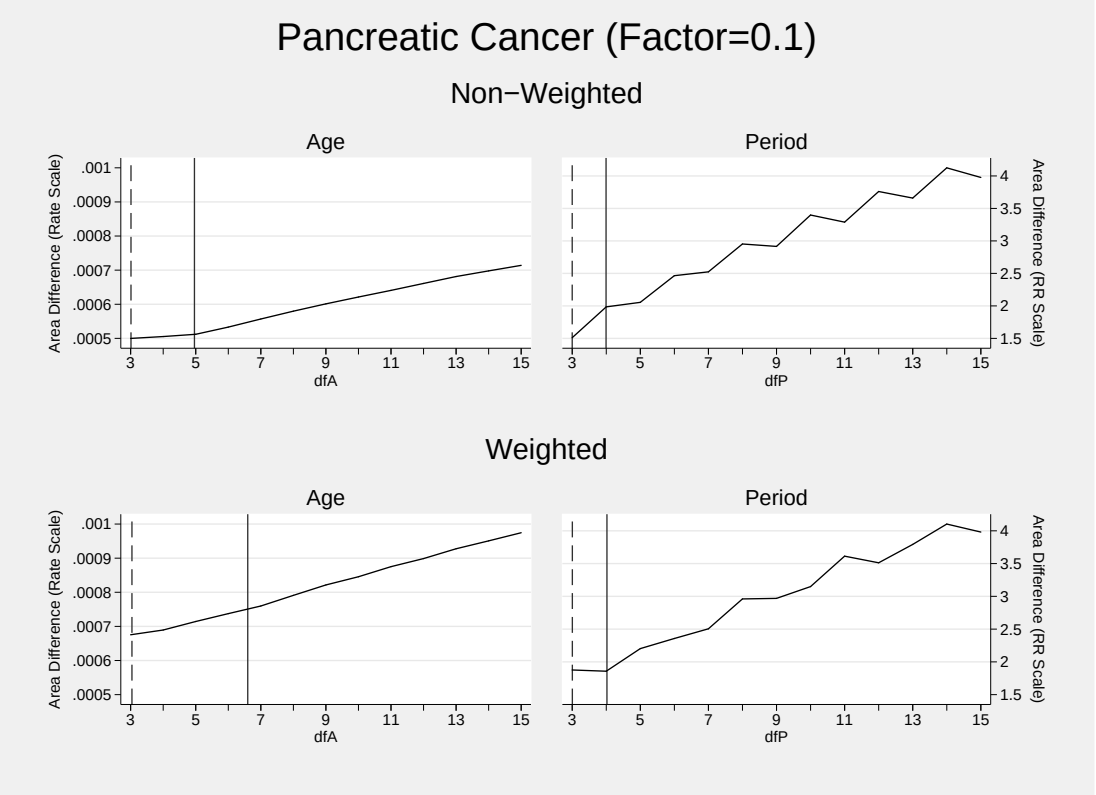

FIGURE 3.17. Results of the full simulation for pancreatic cancer (Factor=0.1).

of the 500 simulated datasets. In the case of the equally-spaced knot placement, it is clear that an increase in the population size reduces the likelihood of local features in the simulated datasets. This leads to the flatter appearance of the area difference curve after the minimum has been achieved.

Figure 3.19 includes the histograms that illustrate the difference between the number of knots selected by the information criteria compared to the "optimal" number of knots for each of the 500 simulations. Negative values on these histograms indicate that the information criteria has selected a smaller number of knots than was optimal, whereas a positive value is indicative of overfitting compared to the optimal knot placement. The optimal number of knots was decided by choosing the number of knots that minimized the difference between the true curve, and any of the compared knot placements (from 3 to 15 degrees of freedom). The three scaling factors (1, 0.1, and 10) are compared in the array of histograms alongside the

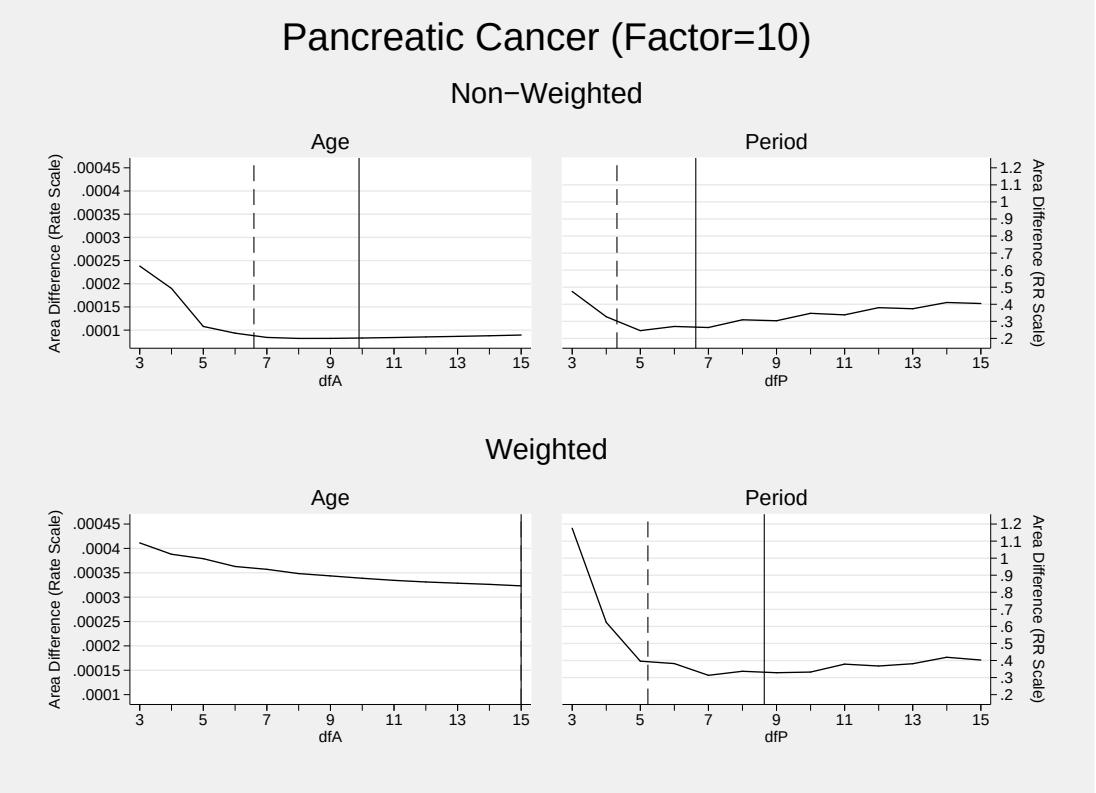

FIGURE 3.18. Results of the full simulation for pancreatic cancer (Factor=10).

comparison between AIC and BIC. Furthermore, both the age and period curves are compared, with age being in the left-hand column, and period the right.

It is clear from Figure 3.19 that the BIC generally selects fewer or the same number of knots as the optimal df and that the criterion rarely selects a model that "overfits". This leads to an almost half-normal shape in some cases, particularly for the histograms relating to the period curve, and when the scaling factor is small (Factor=0.1). On the other hand, the histograms relating to the AIC are generally more evenly distributed, and usually resemble normal curves with a mean of zero. However, there is frequently a significant spread to the distribution, even in the case where the scaling factor is large (Factor=10).

From the figures, it is fairly clear that there can be substantial differences in terms of the number of degrees of freedom selected by the information criteria. One thing that can be drawn from the figures is that it is unlikely that the BIC will select a model that will overfit. This leads to the potential solution that the BIC knot selection can be used as some form of "lower

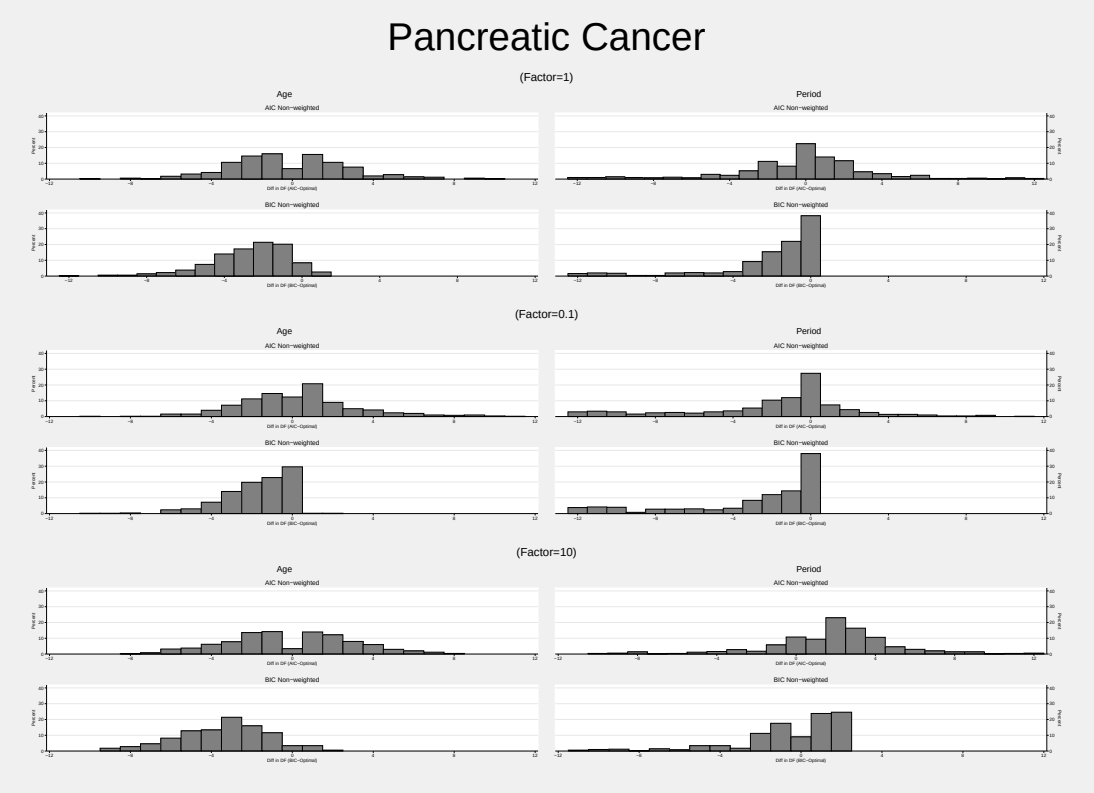

FIGURE 3.19. Histograms showing the difference in df between the selection criteria and the optimal number of knots for pancreatic cancer.

bound" for the number of knots that are selected. The AIC has a tendency to both underfit and overfit, but will generally select a higher degrees of freedom than that selected by the BIC. It is clear that these information criterion can be used as a guide for choosing the appropriate number of knots, but neither can be used as strict rule for selecting the correct number of knots to use.

3.5.2.3. *Simulations based on Hodgkin's Lymphoma*

Figure 3.20 gives the fractional polynomial shapes for age and period that were selected using the Hodgkin's lymphoma cancer dataset. The simulated shape for the age effect has three turning points. The fact that the curve is bimodal over age for Hodgkin's Lymphoma is consistent with results given in the literature in other settings [Glaser, 1991]. The shape for period over time has two turning points.

Figure 3.21 shows the results of the 500 simulations for the dataset based on the Hodgkin's Lymphoma data. It can be seen that in this case, there is a limited range of values that produce

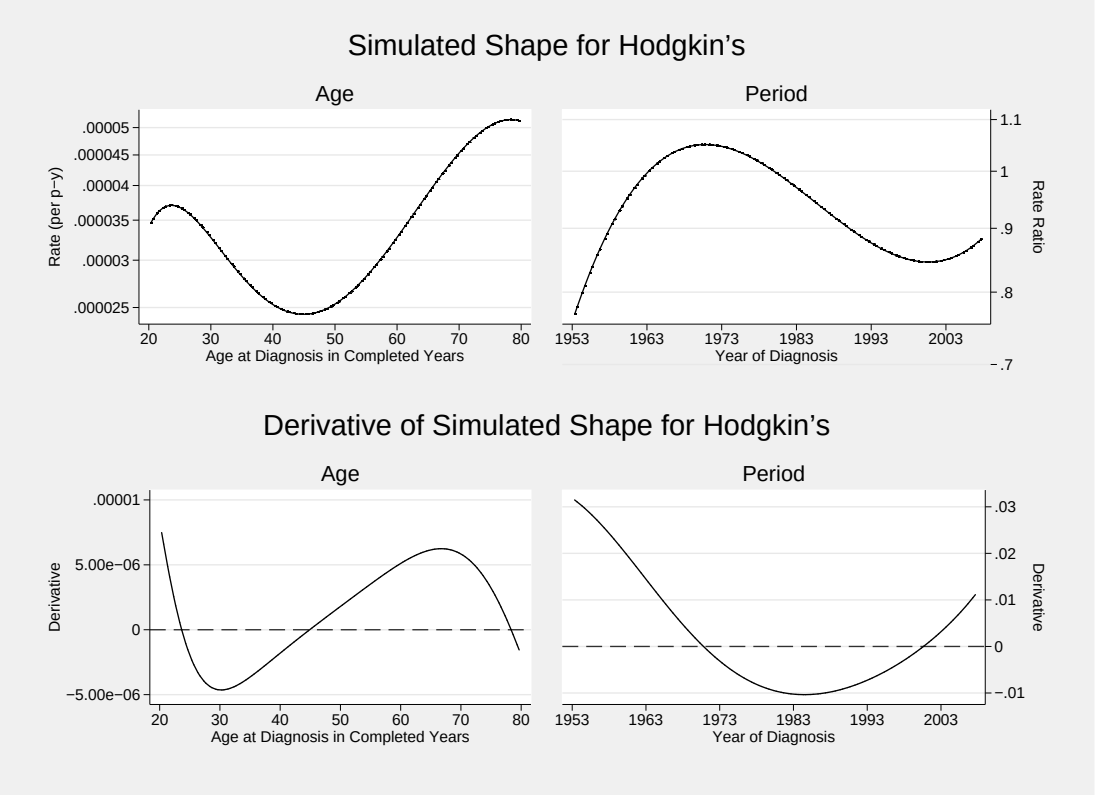

FIGURE 3.20. The true shape generated using fractional polynomials for Hodgkin's lymphoma.

similar results in terms of area difference for the age curve. Before and after this range of values, there is a sharp increase in the area difference indicating that underfitting and overfitting are leading to poorer fits to the "true" shape. For Hodgkin's Lymphoma, the shape for age is similar irrespective of whether a weighted or equally-spaced knot placement is used. This is due to the difference in terms of the distribution of number of cases over the age range that can be seen in Figure 3.20.

The period curves shown in Figure 3.21 highlight that a low degree of freedom is required to capture the shape for Period and that using a higher degree of freedom leads to overfitting. Again, a similar shape is seen for the two choices of knot placement in this example.

Figure 3.22 shows the results of the 500 simulations for Hodgkin's Lymphoma when the cases and population size have been scaled down by a factor of 10. As expected, reducing the amount of information has decreased the degrees of freedom that are selected by the information criterion. From the scale on the figure it is clear that using a reduced amount of information

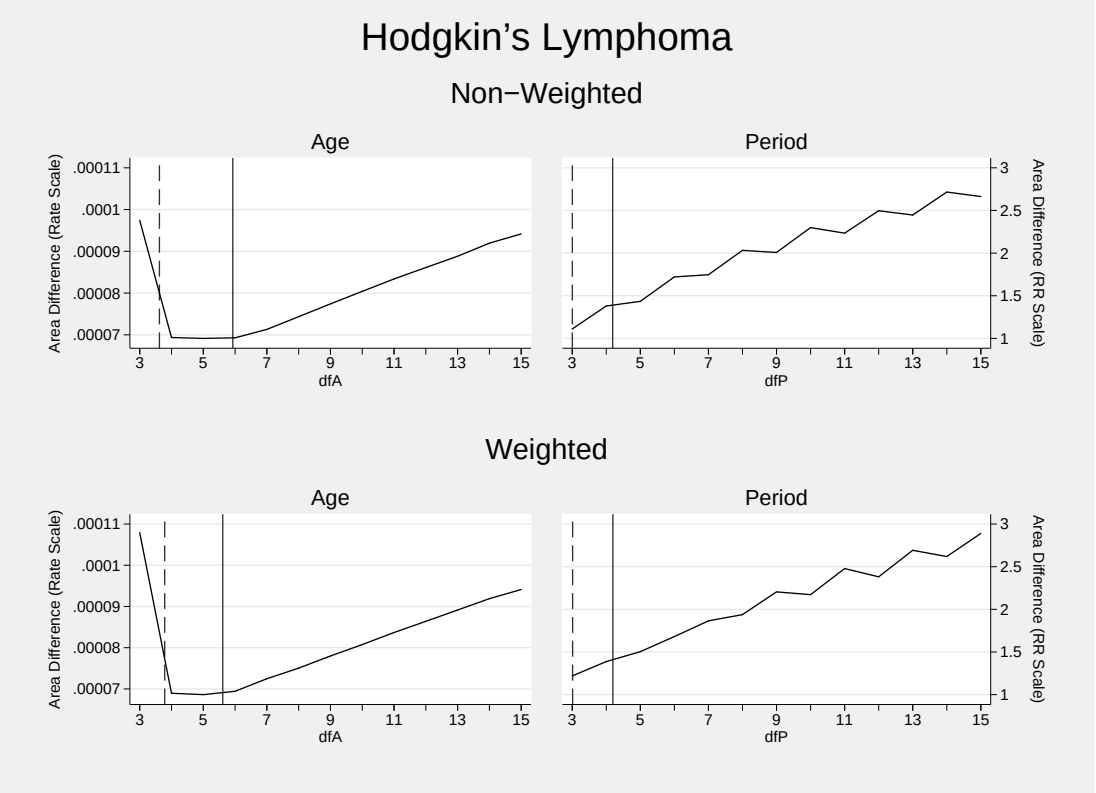

FIGURE 3.21. Results of the full simulation for Hodgkin's lymphoma (Factor=1).

(that is a smaller population size) results in a poorer fit to the "true" shape for both the Age and Period curves.

Figure 3.23 shows the results of the 500 simulations for Hodgkin's Lymphoma when the cases and population size have been scaled up by a factor of 10. Compared to using a lesser amount of information (Factor=0.1,1), the issue of overfitting is lessened when the Factor is increased to 10. This can be seen by the flatness of the curve as the degrees of freedom is increased for the Age curves for each of the knot placements.

Figure 3.24 and Table 3.1 give indications of the difference between AIC and BIC in terms of the number of knots that are selected by each criteria. The figures in Table 3.1 add extra information surrounding the variability in the number of knots selected by the selection criteria for the 500 simulations. Clearly there is less variability for the values for the period curve when compared to the age curve. What is also clear is that the BIC is much less variable in the values

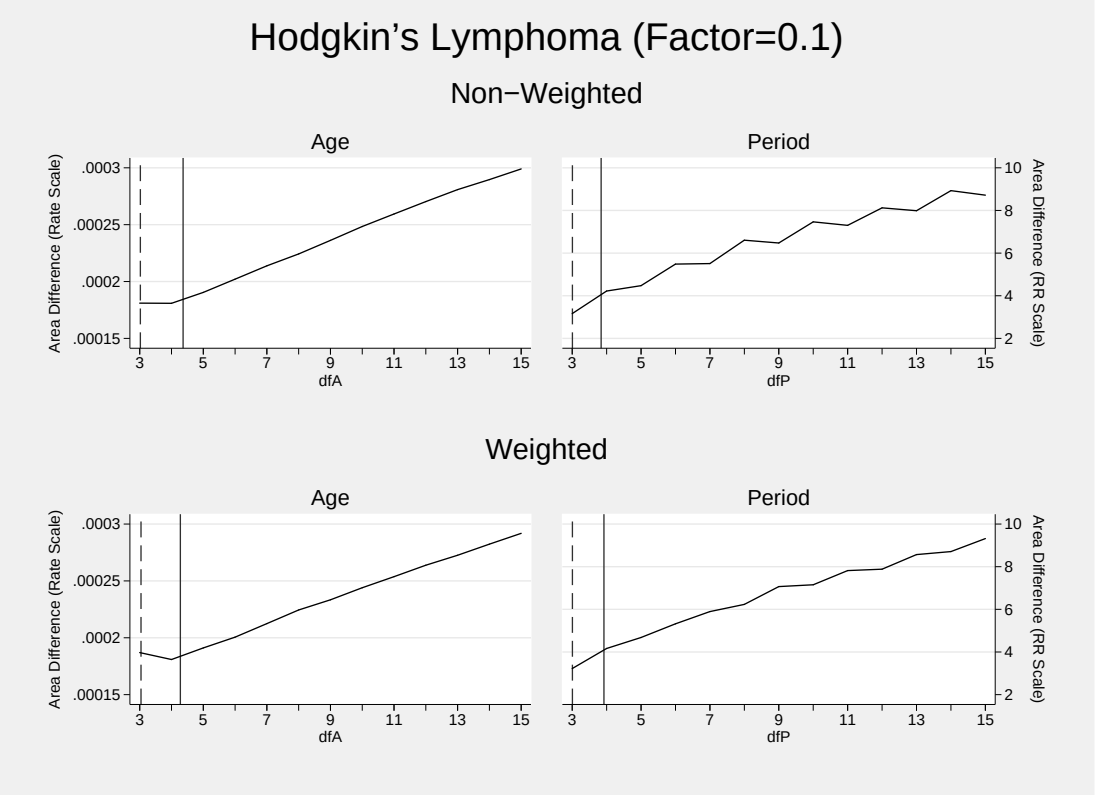

FIGURE 3.22. Results of the full simulation for Hodgkin's lymphoma (Factor=0.1).

Scaling Factor	Curve	Inf. Crit.	Mean	Median	Min	Max	Std. Dev.
Factor=1	Age	AIC	5.928	6	3	14	2.056
		BIC	3.620	4	3	6	0.569
	Period	AIC	4.188	3	3	15	2.139
		BIC	3.004	3	3	4	0.063
Factor=0.1	Age	AIC	4.364	3	3	15	2.351
		BIC	3.020	3	3	4	0.140
	Period	AIC	3.840	3	3	13	1.785
		BIC	3.006	3	3	4	0.077
Factor=10	Age	AIC	9.294	9	5	15	2.320
		BIC	5.970	6	4	9	0.811
	Period	AIC	5.454	5	3	15	2.201
		BIC	3.146	3	3	6	0.466

TABLE 3.1. Comparison of the AIC and BIC values for the various scenarios. Hodgkin's lymphoma.

given than the AIC. This can be seen from the range of values selected by the BIC, as well as from the standard deviation reported in the final column.

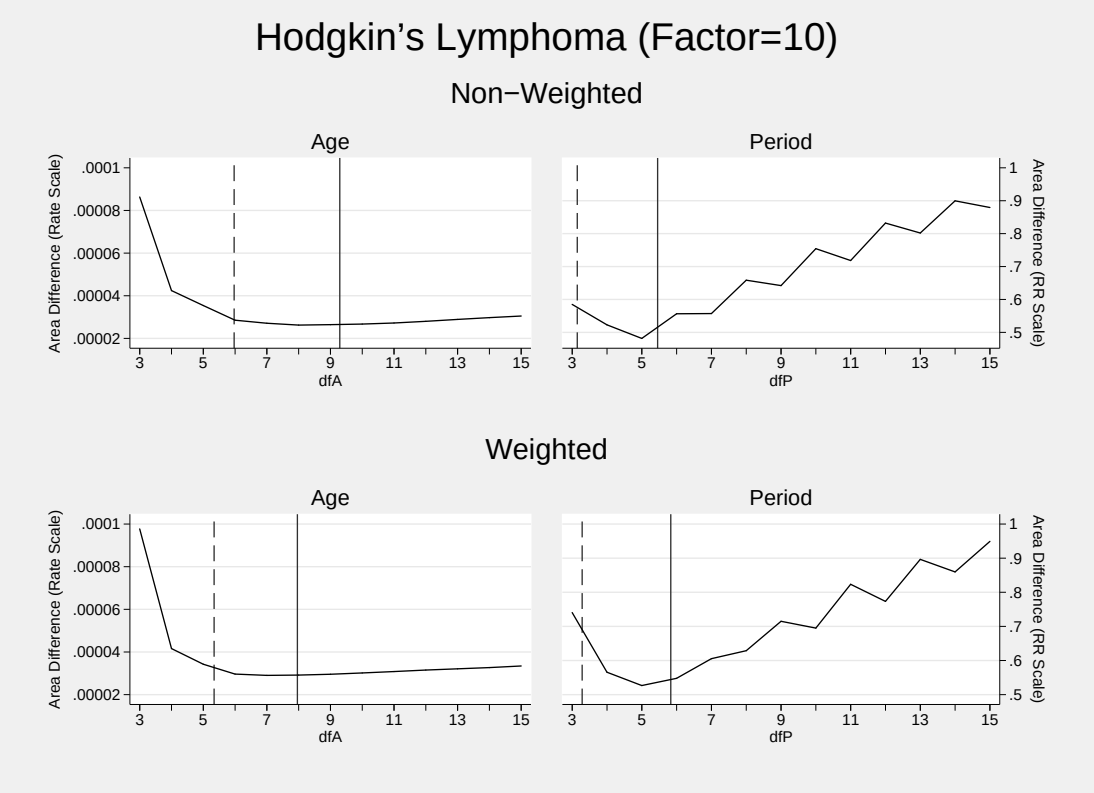

FIGURE 3.23. Results of the full simulation for Hodgkin's lymphoma (Factor=10).

Figure 3.24 compares the degrees of freedom values selected by the AIC to those selected by the BIC. What can be seen is that the AIC value is greater than or equal to the BIC value in every scenario for each of the simulations. This is indicative of the higher penalty that is introduced by the BIC to ensure that the model does not overfit. The information in Figure 3.24 coupled with the information in Table 3.1 can be used to understand the difference that the additional penalty term has in terms of the degrees of freedom selected.

3.5.3. Explaining the zig-zag

Looking at the results for the simulations (Figures 3.10, 3.16, and 3.21), there was an unexpected jagged shape to the graphical display of the main results for each of the scenarios for the Period curve (that is, the curve displaying the area difference between the fitted function and the true simulation shape for a range of df values). This shape seemed to suggest a case of a difference in terms of fit between knots placed using an even number of degrees of freedom compared to an odd number.

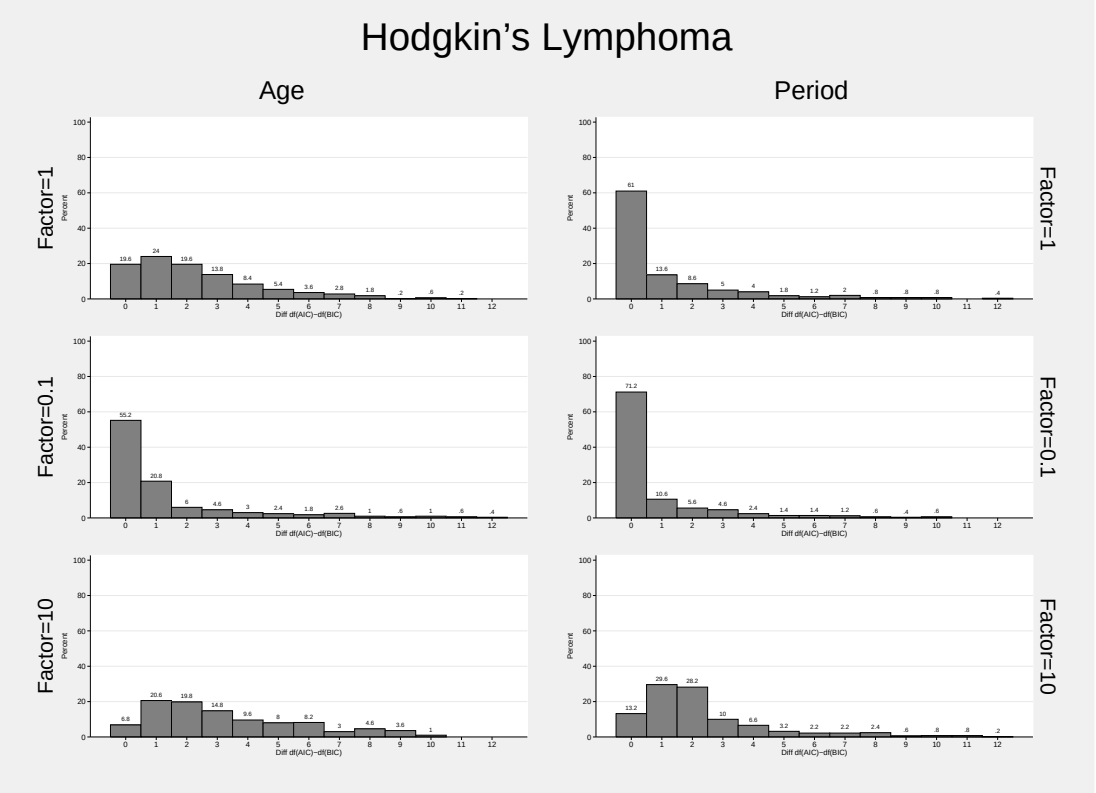

FIGURE 3.24. Comparing the difference in df selected by the AIC and BIC. Hodgkin's lymphoma.

Figure 3.25 shows the results of a single simulation for the period curve for lung cancer with 3 fitted functions having different degrees of freedom. The knot placements are also shown on the graph. The "truth" is shown in orange and has a peak at around 1975. In this single simulation, there appears to be a suggestion that there is a dip at the same point, which is picked up by the fitted functions with an even degrees of freedom, but is "missed" when the degrees of freedom is set to 9 for the period curve. The knot placements for the even degrees of freedom intersect the period of the "dip", but the knot placement for the odd degrees of freedom falls a substantial distance from the "dip" on either side and consequently the fitted function averages over this period rather than being flexible enough to pick up the feature. This results in the fitted line for 9 degrees of freedom being closer to the "truth" than both the fitted function with 8 and 10 degrees of freedom. This leads to the jagged shape observed for the area difference curves shown as the main results of the full simulation.

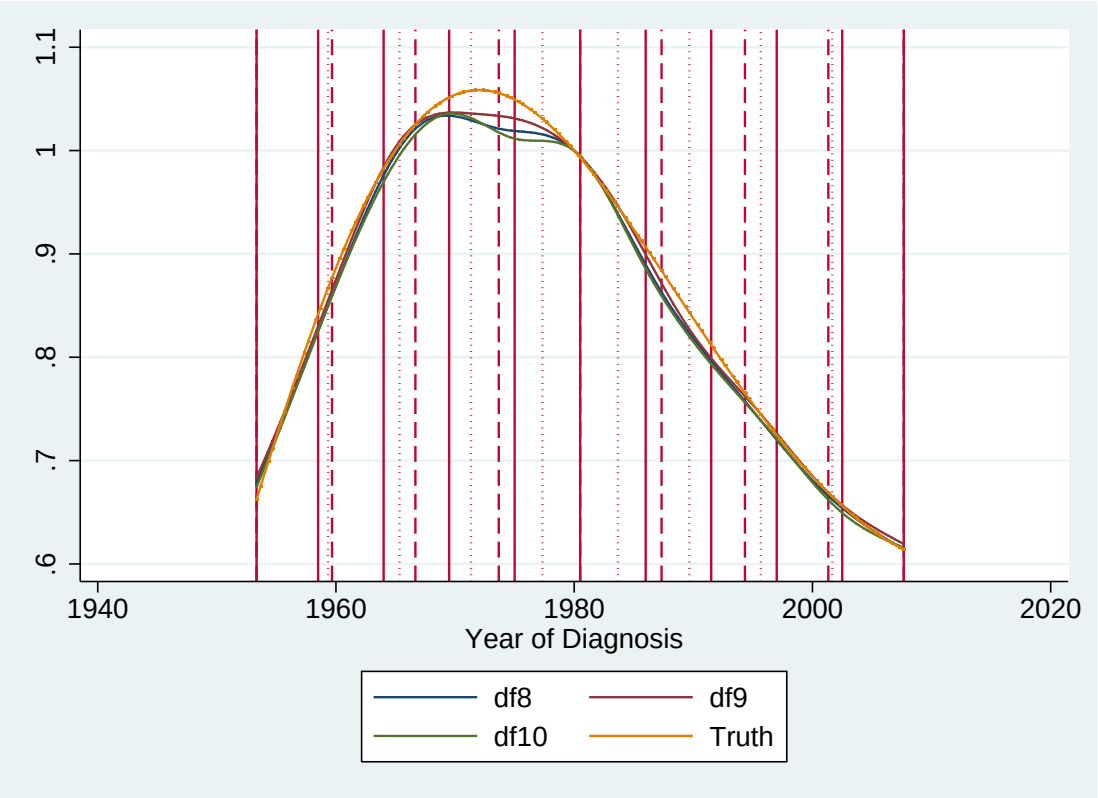

FIGURE 3.25. Explaining the zig-zag. Single simulation for the period curve for lung cancer.

Figure 3.26 shows the knot placements that are employed for the equally spaced centiles of period for the entire range of degrees of freedom (3 to 15). The pattern that is generated for both the odd and even degrees of freedom show different clusterings of the knot placements; dependent on whether an odd or even number of degrees of freedom are used (this is particularly the case for the higher degrees of freedom). This is a natural occurrence on the basis that each of the even numbers is divisible by 2, and some of the odd numbers also share common factors. It is also clear that the odd degrees of freedom do not come particularly close to placing a knot near to the 50^{th} centile of the distribution, even though upto 15 degrees of freedom are included in the illustration. This figure corroborates the finding from the previous illustration, in that the knot placements for the odd and even degrees of freedom can be fundamentally different at some points across the timescale. If, as is the case for lung cancer, this coincides with a region where there are more cases, this may well lead to the jagged shape for the area difference curves shown as the main results of the full simulation for the period curves.

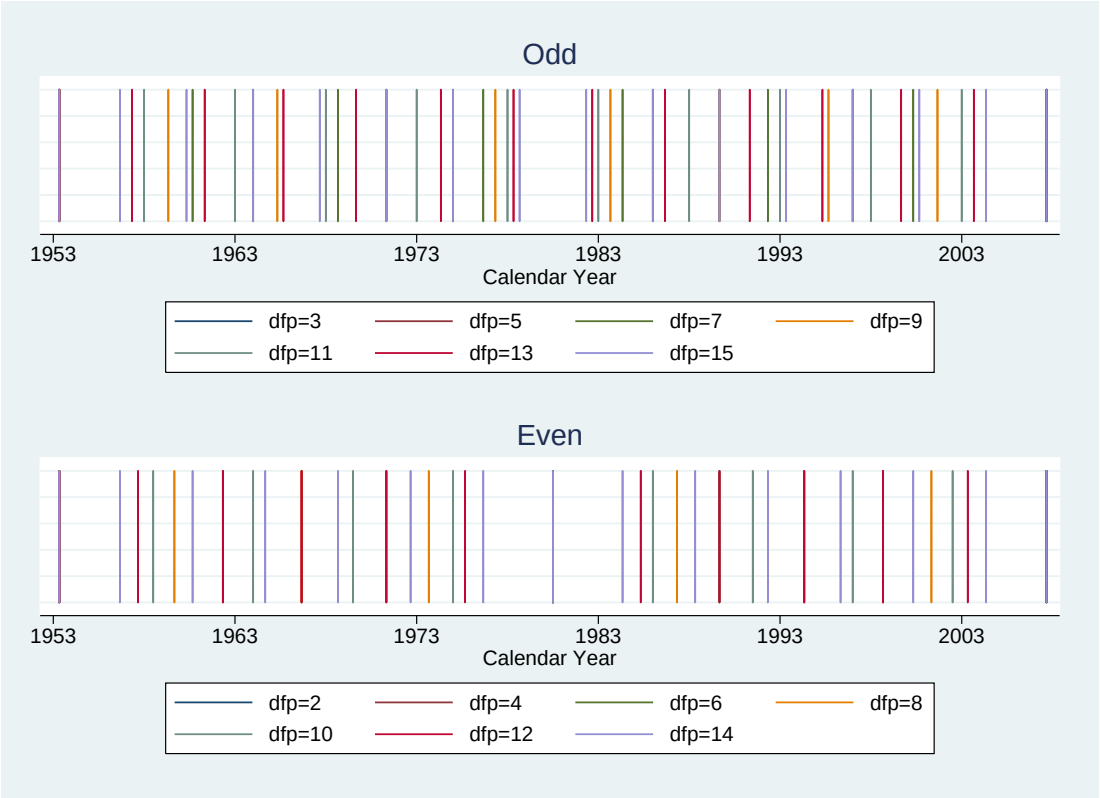

FIGURE 3.26. Comparing the knot placements for odd and even degrees of freedom.

In order to further corroborate the evidence seen in single runs of the simulations, a second simulation was undertaken to investigate whether the placement of the knots for the odd and even degrees of freedom was causing the jagged shape for the area difference curves.

To make a comparison, the standard equal knot placements were compared to a knot placement that had been randomised in a way that did not lead to extreme knot placements being compared. This was acheived by randomly selecting the knots from a given range either side of the usual knot placement whilst still appropriately using the entire range of the potential centile values. For example, for 5 degrees of freedom the internal knots are usually placed at the $20^{\text{th}}, 40^{\text{th}}, 60^{\text{th}}$ and 80^{th} centiles of the variable for period. This was compared to a "random" knot placement where the centile values for the 4 internal knots were taken randomly from the centile ranges of $11^{\text{th}} - 30^{\text{th}}$, $31^{\text{st}} - 50^{\text{th}}$, $51^{\text{st}} - 70^{\text{th}}$ and $71^{\text{st}} - 90^{\text{th}}$. This same pattern was undertaken for 200 simulations for each of the degrees of freedom values from 3 through to 15. The resulting output of the comparison can be seen in Figure 3.27. It is clear that the "jagged"

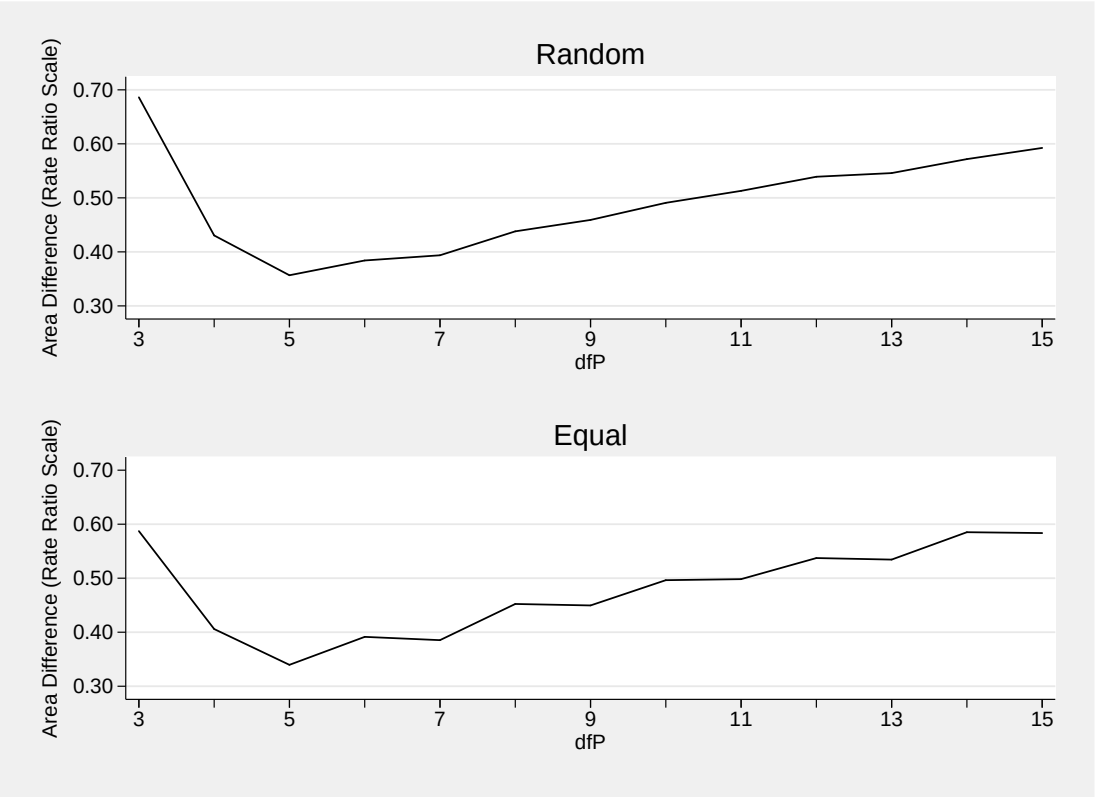

FIGURE 3.27. Results of the simulation using a standard compared to a "random" knot placement.

appearance of the line using the equal knot placements is dramatically reduced when using the randomisation procedure for the knot placement that is described above. This lends further evidence to the hypothesis that the unexpected results of the full simulation are caused by how the knots are placed using the centiles of the distribution for the period term.

An interesting feature of Figure 3.27 is the fact that the average area difference for 3 degrees of freedom is worse for the random knot placement compared to the equal knot placement. This is due to the fact that inefficient knot placements are allowed by the random knot placement technique. That is, the knots could be placed quite closely together or quite extreme distances apart compared to the equal knot placement. This effect is lessened as the degrees of freedom are increased as the window in which any one knot can be placed is reduced in width due to the method used for "randomly" placing the knots.

3.6. Discussion

The issue of knot placement is commonly raised when spline functions are used as part of any analysis. The criticisms of the spline functions being overly sensitive to the number and placement of the knots are usually exaggerated. The shape of the spline functions are very much dictated by the available data and the decision on how many knots to choose can often be determined by the setting. If a smoothed function to describe "noisy" data is required then fewer knots should be chosen and the BIC tends to provide an appropriate guide for a "simpler" shape due to its higher penalty function. If local minima and maxima are expected and a fairly complicated shape is required then more knots will be required in order to capture this complexity. The AIC tends to select a higher degree of freedom than the BIC and may well be a starting point for this type of analysis. Irrespective of the setting, it is good practice to perform a simple sensitivity analysis to check how sensitive the fitted function is to the number and placement of the knots; this is a simple extension to model selection techniques and the "model-building" process.

In the case of using restricted cubic splines for age-period-cohort models, the simulation carried out in this chapter highlights that as long as "enough" knots are chosen, it does not matter overly if a few too many knots are selected. A few knots either side of the "best" number of knots for any given fit, generally will not alter the fit dramatically. The information criterion can have a role to play in determining the selection for the number of knots. The AIC will tend to select more knots than the BIC, but they generally agree to within 1 or 2 knots in most cases.

The placement of the knots should also be considered alongside the selection of how many knots to choose. However, the analyses in this chapter, and from experience with analysing the Finnish data suggest that placing the knots at equal centiles of the data (dependent on the number of knots chosen) is generally sufficient as a strategy. It may be possible to achieve the same fit with fewer knots if they are placed carefully. However, in the population-based registry data, there is usually sufficient data that it matters little if a few extra degrees of freedom are used unnecessarily. In other settings, with smaller datasets, knot placement may well be a bigger issue and require further effort. The weighted knot placement is possible in this setting, and may well save a few degrees of freedom. However, this is at the cost of potentially having

a poor fit for the youngest ages in the age and cohort terms, which could lead to important features being overlooked; the reduced level of information does not necessarily mean that linearity should be enforced.

For the simulation process, fractional polynomials were used to generate the true shape. This choice led to the curves for the age effect to have an early turning point for each of the sites; which is where there are fewer cases in order to appropriately capture the shape. Consequently, the weighted knot placement approach performed poorly for the age curve for the simulations. This may not give a true reflection of what is likely to be of interest in practice and, therefore, the weighted knot placement could still be considered as a viable approach in real applications.

A further comparison that could have been undertaken as part of the simulation process would have been to compare the fitted functions for the various spline models to the fitted function from the fractional polynomial model that was used to generate the data. This would have given a further insight into how well the spline models fit to the data. This comparison would have shown how many knots would have been required in order to get a sufficiently close fit to the data generating model in each of the simulated datasets.

In this chapter, the size of the datasets was also varied by using a simple multiplication factor to create a larger and smaller population size than Finland for the specific cancer sites. It is clear from the results of this investigation that the size of the dataset can have an influence on the complexity of the shape that is selected by the information criteria. However, even though a larger degree of freedom may well be selected with an increase in the size of the dataset, the shape of the curve is more stable because of the extra information to dictate the shape. This can be seen by the histograms that show the number of turning points. When the Factor is set to 10, even with 15 degrees of freedom, the number of turning points is the same as the truth in 100% of the cases. The random noise around the true shape is reduced due to the increase in information. If age-period-cohort models are applied to regional cancer registry data, it is clear that more care should be taken when deciding upon the number of knots that are used.

The differences that have been seen in this chapter between different knot placements have usually been small in absolute terms. For example, the differences seen in Figures 3.5 and 3.6 show that even with a large difference in the number of degrees of freedom selected, a similar fit is often achieved. Provided that a sensitivity analysis assessing the number of knots is carried

out as part of any study, it seems unlikely that knot selection will cause any interpretational issues for age-period-cohort models using restricted cubic splines.

CHAPTER 4

Incidence Projections

4.1. Chapter Outline

In this chapter a novel approach to incidence projection is discussed based on the age-period-cohort modelling approach using restricted cubic splines that was introduced in Chapter 2. An overview of previous methods for incidence projection will be given, and then the most commonly applied method will be compared to the newly suggested approach. The issues over knot selection that were studied in the previous chapter are also reconsidered in light of the modification required in order to make the projections. Most of the work detailed in this chapter is covered in a paper that is currently undergoing peer-review [Rutherford et al., 2011b]. The draft of the paper is given in Appendix III.

4.2. Literature Review

Methods for projecting incidence have been given a lot of attention in the cancer epidemiology literature. Motivation for making projections is given by the fact that resources need to be appropriately allocated in the future for cancer care [Theisen, 2003]. In a review article discussing the prediction of future burden, Bray and Møller discuss the fundamentals of the approaches and provide a good review of recent methodology [Bray and Møller, 2006].

Early literature on projecting from age-period-cohort models covered the issue of ensuring that the projections made were independent of the chosen parameterisation for the model [Osmond, 1985]. There have been numerous methods proposed for making projections from age-period-cohort models [Clements et al., 2005; Bray and Møller, 2006], and there are a number of examples of projections being made in practice [Rostgaard et al., 2001; Agha et al., 2006; Cleries et al., 2009]. An evaluation of using Age-Period-Cohort (APC) models for projection using a variety of methods has been carried out using data from the Nordic countries [Møller et al., 2003]. Some of the methods included are based on recommendations from an earlier

evaluation using Nordic data [Engeland et al., 1993]. From this extensive comparison, software has been developed for the preferred methods [Engholm et al., 2009].

Bayesian analyses have become popular for age-period-cohort modelling [Nakamura, 1986; Berzuini and Clayton, 1994; Besag et al., 1995] and there have been extensions of the Bayesian approaches in order to make projections [Bashir and Estève, 2001; Bray et al., 2001; Knorr-Held and Rainer, 2001; Wong et al., 2007]. These approaches use a variety of smoothing priors in order to make sensible projections from the models.

Europe-wide projections of incidence and mortality have been made [Ferlay et al., 2007], and recent projections have also been made in the UK [Møller et al., 2007]. In addition, there have been large-scale studies using SEER data in the US [Smith et al., 2009]. Arguments have also been made that care must be taken to consider the differing sub-populations at risk when making projections considering that incidence can differ substantially across race [Wong et al., 2007].

There are a number of examples where particular care has been taken to understand fully the impact of projections for a particular cancer site, rather than using an overall approach to many sites [de Vries et al., 2005; Wong et al., 2007; Beelte et al., 2008]. The merits of this approach will be discussed in the following chapter.

A number of approaches have been motivated using mortality rates rather than incidence [Knorr-Held and Rainer, 2001; Olsen et al., 2008; Shibuya et al., 2005]. However, the methods can easily be applied to incidence data instead. The mortality data can be expressed in exactly the same format as the incidence data, with new cases of cancer being replaced by cancer-specific deaths.

A discussion has been given in a Bayesian setting of the effect of excluding the youngest age groups when making incidence projections [Baker and Bray, 2005]. The authors argue that improved projections can be made by including the youngest age-groups despite the lack of data. However, a reply to this article [Clements et al., 2006] argues that the priors employed in the Baker and Bray approach lead to wide credible intervals for the projections, and argue that other approaches perform more favourably in this respect [Clements et al., 2005; Møller et al., 2003].

Simple methods for incidence projection have been made using the Finnish registry data in the past [Hakulinen et al., 1986; Dyba and Hakulinen, 2000, 2008]. These methods use simple interpolation techniques for the incidence rates in order to make the projections and they assess whether or not reliable projections are obtained. The methods proposed allow the age-period specific number of observed cases to follow a Poisson distribution, and they compare linear projections of the age-specific incidence rates on both a linear and log scale. These methods are similar in principle to the projections that would be obtained from an age-period model, that ignores the effect of cohort.

4.3. Introduction

As discussed in Chapter 2, an age-period-cohort model provides a modelling tool that can be used to describe the rate of either the incidence or mortality of a given disease. In order to obtain an estimate of future cancer burden, it is necessary to appropriately project estimates of cancer incidence. A detailed evaluation of using Age-Period-Cohort (APC) models for projection has been carried out using data from the Nordic countries [Møller et al., 2003]. The authors (Møller *et al.*) compared 15 different methods and drew general conclusions on the best strategy for obtaining appropriate future predictions of incidence. They found that the APC models that used a standard log link function tended to give an overestimate of future incidence rates, and proposed an alternative power link function in its place. They also argued that the linear term used for the projections should be tempered ("dampened") for longer term projections in order to avoid over-estimation of the true rates. The models compared by Møller *et al.* [2003] were largely based on coarsely grouped data and mostly relied on factor models to make the predictions into the future. It could well be more appropriate to use finely-split data and use a smoothing technique such as cubic splines in order to make estimates of the incidence [Heuer, 1997; Carstensen, 2007] and then project into the future using these models.

As introduced in Chapter 2, smoothed estimates can be obtained from using restricted cubic splines for the three terms; age, period and cohort. The benefit of choosing to use restricted cubic splines as the smoothing method for the finely-split data is the restriction that is then enforced beyond the boundary knots that are specified. The fitted function is forced to be linear beyond the final boundary knot (as well as prior to the first boundary knot). This restriction

can then be utilised for making a linear prediction beyond the realms of the data, whilst still being dictated by the shape of the data towards the end of the observation period.

In order to validate the new approach to making projections, the method is applied to 4 of the commonest cancer sites (breast, lung, colon and pancreas) from data made available by the Finnish Cancer Registry [Finnish Cancer Registry]. Comparisons are drawn between the methods that use the linear constraint of the restricted cubic splines and a more standard method for projecting incidence from age-period-cohort models.

4.4. Description of the Data

To illustrate the methods and assess the effectiveness of the new technique, the methods are applied to four common cancer sites. The chosen sites are breast cancer (females only) and lung, colon and pancreatic cancer (analysed separately for each gender). The cancer registry data used for the analyses contains incidence data from 1953 until the end of 2007. In order to assess the projection methods, it is necessary to use a retrospective analysis. This means that the methods of projection are directly compared to what actually happened in the succeeding years by making the projections from an earlier point in time. On the basis that both short, and long-term projections are of interest for health planning authorities, 10 and 20 year projections were undertaken. In order to make 20 year projections, it is necessary to use observed data that ended at, or prior to, the end of 1987.

Not only are the incidents of cancer needed from the cancer registry, the population size is also required for the denominator in the incidence rate estimation. In the following analyses, the known population size is used when projecting into the future rather than projecting both incidence and the population size. Population statistics are made available by Statistics Finland [Statistics Finland] for the Finnish population split by age, and calendar year. Carstensen [Carstensen, 2007] outlines a method that makes appropriate calculations using formulae suggested by Sverdrup [Sverdrup, 1967] to estimate the risk-time for the triangular subsets of the Lexis diagram; split by age, period and cohort (this is detailed in Section 2.4.1). The formulae suggested appropriately account for the population size in the relevant years, and age-groups. This means the data are split more finely than the yearly interval splits, and requires that appropriate averages are taken for the triangular subsets of the Lexis diagram to be values of

age, period and cohort. This data format will be used for the models using the restricted cubic splines to smooth the effects.

4.5. Methods

4.5.1. Age-period-cohort models

As introduced in Chapter 2 (Section 2.3), the general form of the age-period-cohort model (with $\mathbf{a} = \mathbf{p} - \mathbf{c}$) can be given as:

$$\ln\{\lambda(\mathbf{a}, \mathbf{p})\} = f(\mathbf{a}) + g(\mathbf{p}) + h(\mathbf{c}), \tag{4.1}$$

where f, g and h are functions, and $\mathbf{a}$, $\mathbf{p}$ and $\mathbf{c}$ are the values of age, period and cohort respectively.

The model used by Møller et al. [Møller et al., 2003] is expressed as a factor model for age, period and cohort, with the levels of the factor expressed by the subscripts:

$$\ln\{R_{ap}\} = A_a + Dp + P_p + C_c, \tag{4.2}$$

where R_{ap} is the incidence rate in age group a and calendar period p, D is the drift parameter, A_a is the age component for age group a, P_p is the non-linear period component of period p, and C_c is the non-linear cohort component of cohort c [Møller et al., 2003].

Due to the identifiability issue (Section 2.6), it is widely accepted that only some of the components of the model expressed in equation (4.1) can be uniquely determined [Holford, 1983; Clayton and Schifflers, 1987b]. Carstensen [Carstensen, 2007] discusses, at length, the principles of parameterisation that are desirable in order to apply appropriate and meaningful constraints to extract identifiable quantities from the models. In the factor model described above in equation (4.2), the drift parameter has been extracted, and consequently, the factor terms given for period (P_p) and cohort (C_c) have been "de-trended". That is, they give the non-linear effect of period and cohort respectively. Removing the linear trend via the matrix transformations described in Section 2.8.2 is one method for overcoming the identifiability issue, and fitting the model described in Equation (4.2). This could equally be achieved by dropping one factor level each from the period and cohort terms. The underlying models that are fitted

are equivalent; they are different parameterisations of the same model, and the estimates for the overall rate produced from the models are identical.

However, when projecting the incidence rate, is it vital that the method employed to make the projection is independent of the parameterisation that is chosen [Osmond, 1985]. The projections that are made from the methods that are proposed in this chapter do not depend on the chosen parameterisation of the model.

4.5.2. Using the Linear Constraint of Restricted Cubic Splines for Projection

Instead of using a step function as the functional form in equation (4.1), it is possible to consider using a smoothing function for each of the terms, such as splines (as introduced in Section 2.8.1).

Restricted cubic splines are used for each of the three components of the age-period-cohort models (that is, as the functions defined in equation (4.1)). Due to the identifiability issue that is associated with APC models, the spline functions must be subject to a further constraint [Carstensen, 2007; Rutherford et al., 2010]. The constraint that is usually applied extracts the combined linear trend (the drift) from the period and cohort terms, which effectively fixes the slope of both curves. The linear drift is then attributed to either of the two terms (period or cohort); leaving the other term to have no overall trend (see Section 2.8.2).

The period and cohort terms both have spline functions to describe their shape. On the basis that it is usual to make simple assumptions when projecting, it is possible to make projections into the future from both of these terms by extending them linearly (see Figure 4.1). Due to the restriction of the cubic splines, this is easy to achieve as the function fitted by the splines for both period and cohort are forced to be linear beyond the boundary knot (which would usually be placed at the last observed data point). Using the restriction of the restricted cubic splines to make the projections means that the linear trend that is projected will automatically be determined by the latter part of the data. In order to stabilise the estimates, it is also possible to bring the final boundary knot within the range of the observed data and enforce a linear trend to occur from an earlier point in time. In the analyses that will be carried out in this chapter, the boundary knot for period and cohort is moved 10 years into the range of the respective observed data for both terms (this knot placement is shown by the dashed vertical

lines in Figure 4.1). A sensitivity analysis is carried out in the latter part of Section 4.6.1 to assess the impact of selecting 10 years as the proposed length.

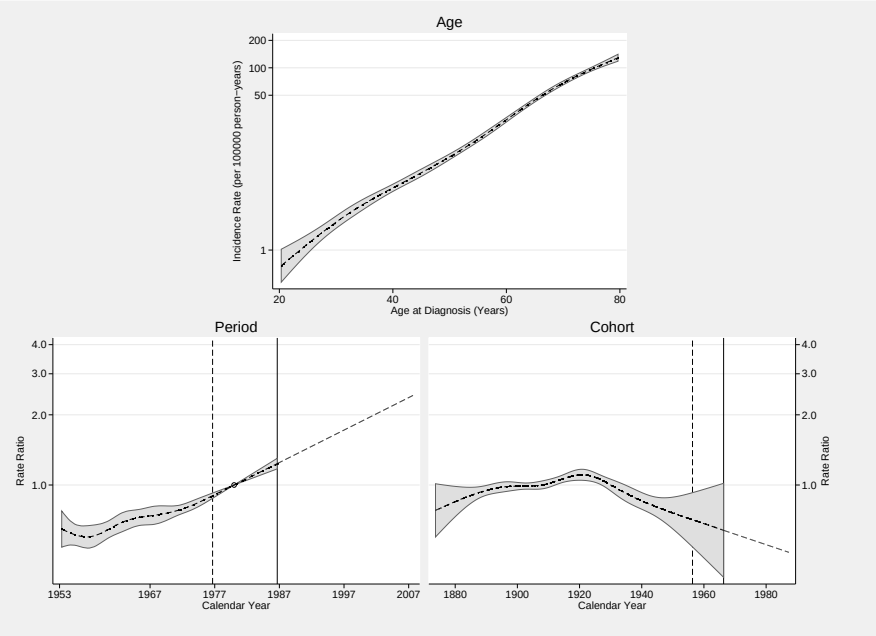

FIGURE 4.1. Example of the graphical representation of the age-period-cohort model using restricted cubic splines. The data used are for the incidence of Finnish colon cancer for males. The drift term is attributed to the period curve, and the age curves are the fitted rates in the reference period (1980; indicated by the hollow circle).

4.5.3. Projecting the Full Drift

A standard method for projecting incidence is to project the drift (Dp) term into the future. This method is advocated by Møller et al. [Møller et al., 2003] as the simplest method of using the APC models to provide a linear prediction into the future. The model used for this method is given in the factor model described in Equation (4.2). The factor model advocated uses data for age and period in five-year intervals.

A similar approach can be carried out for the APC models using restricted cubic splines because the linear drift is extracted as part of the model constraints. The drift term can be projected by using the estimate of D to make future estimates for the period term, and consequently calculate the estimate of age-specific incidence for future time-points. The advantage of

using the spline model is that yearly data can be used for age and period whilst still obtaining a smooth estimate of the functions.

Comparisons will be given for the obtained estimates of future incidence by projecting the drift term using both the factor model and the model that uses restricted cubic splines. The difference between the two methods is simply how smoothly the age, period and cohort effects are estimated from the given data, and the difference that makes to the estimate of the drift. A further difference relates to the consequent smoothness of the projections made from the two methods. The fact that the data are split into five-year time periods for the factor model means that the predictions are really only valid as an average level over a given five-year period. If yearly predictions are required by health-policy planners then the methods using the more finely split data are more appropriate. The splines make an appropriate smoothing of this finely-split data, with the complexity determined by the number of knots chosen.

4.5.4. Altering the Link Function

The standard link function used for age-period-cohort models is the log link function. It has been suggested [Møller et al., 2003; Engeland et al., 1993] that the exponential growth that this introduces for the projections leads to an overestimation in the projected incidence, particularly for long-term follow-up. These two papers propose a power link function (with a power of $\frac{1}{5}$; often referred to as the Power 5 model) to level of the exponential growth. To be consistent with the literature, this model will be referred to as the Power 5 model in this chapter.

The model with the alternative link function can be described as;

$$\{R_{ap}\}^{\frac{1}{5}} = f(a) + g(p) + h(c). \tag{4.3}$$

The power link function can be used for each of the desribed measures and a comparison between the two link functions will be made for each of the approaches.

4.5.5. Compared Methods

To summarise, the methods that will be compared are given in list format below.

(1) Methods using the Log Link.
   (a) Using the linear restriction of the cubic splines (new method).
   (b) Projecting the drift from the spline model.

(c) Projecting the drift from the factor model (5-year data).

(2) Methods using the Power (5) Link.

(a) Using the linear restriction of the cubic splines (new method).

(b) Projecting the drift from the spline model.

(c) Projecting the drift from the factor model (5-year data).

4.5.6. Simple Description of the Methods

The newly proposed method using the linear constraint of the cubic splines (method (a) in the list given in section 4.5.5) essentially uses linear projections of the spline terms for the Period and Cohort terms (with either one of these including the drift). Moving the boundary knot within the range of the data (10 years) for both the period and cohort terms is an attempt to give more stable estimates from this approach. The linear constraint of the cubic splines means that it is easy to extend the fitted functions for period and cohort into the future. The age effects are assumed to be the same for the projections as they were for the fitted data.

The methods that project the linear drift (method (c) in the list given in section 4.5.5) do so by extracting the linear component from the Period and Cohort terms, and then fitting the non-linear Period and Cohort terms as part of the model (as shown in Equation (4.2)). A similar process can be achieved in the spline modelling framework (method (b) in the list given in section 4.5.5), with spline functions used to express the non-linear period and cohort terms. Again in this setting, we assume that the age effects are the same for the projections as they were for the fitted data. The linear drift over the entire range of the data is then projected in order to estimate the rates in the future. The future non-linear components are set equal to the last estimated effect in the model [Møller et al., 2003].

Each of the methods above can also be applied with any given link function as all of the methods can be expressed as GLMs. Therefore, further comparisons are made by varying the link function from the standard log link function to the power link function with the exponent of $\frac{1}{5}$. It is expected that, particularly for longer-term predictions, the power link function will give better predictions, as the log link function tends to give over-estimates of the true rate [Engeland et al., 1993].

Figure 4.2 illustrates the expected benefit of using the linear constraint method to obtain a more recent "trend". The straight line indicating the "full drift" takes the linear shape over

the entire range of follow-up, which leads to the over-estimation in this illustration due to the tempering of the incidence rate. The spline restriction method in this case uses the last 10 years to define the linear projection beyond the end of the observed data leading to a projection that is closer to the true rate.

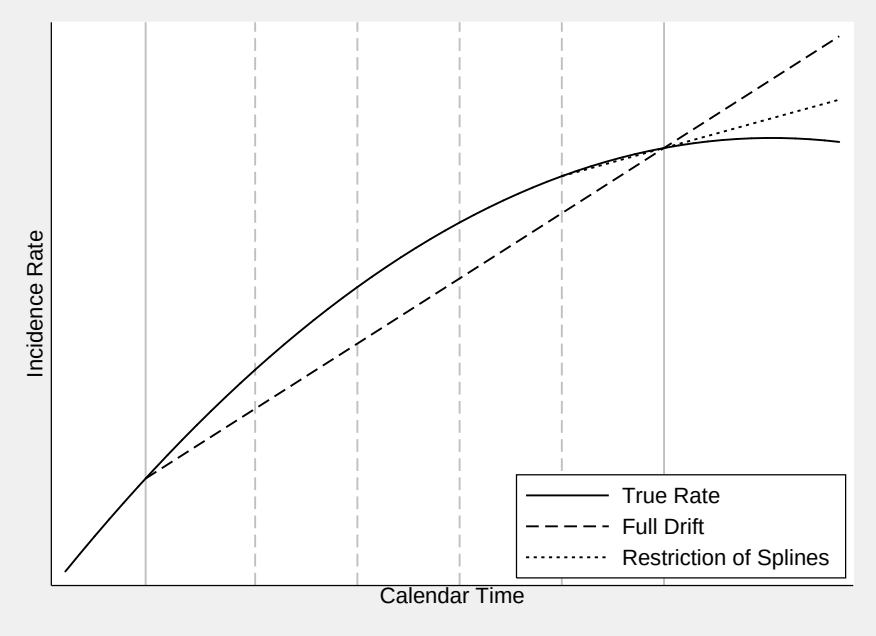

FIGURE 4.2. Comparison of short versus long-term predictions.

By using splines to model the shapes of age, period and cohort rather than factor models, a more realistic shape for the incidence curves can be produced. For each of the spline models that are compared in the results section, 5 degrees of freedom are used for the period term, whereas 8 degrees of freedom will be used for the age and cohort terms. A sensitivity analysis into the choice of the number of knots is given in Section 4.6.1.

4.5.7. Method of Comparison

Long (20-year) and short-term (10-year) predictions are to be compared for each of the methods. The approach taken for the comparison is similar to that undertaken in the previous comparisons made for the Nordic countries [Møller et al., 2003]. The analysis was conducted for all ages combined, and the absolute value of the total relative difference between the observed and predicted number of cases ($|\text{observed} - \text{predicted}| * 100\%/\text{observed}$) was used as the method of comparison. The comparison undertaken in previous analyses uses the total number

of predicted cases over a five-year period, and compares that to the total number of observed cases over the same five-year period. This method of comparison nullifies the benefit of the approaches using splines for the yearly split data. In order to do a fairer comparison, we can use the same measure of relative difference, but apply this measure to each year of the five-year prediction window. It is then possible to report the mean of these five numbers as a measure of the effectiveness of each prediction technique over the five-year period.

The long term projections were made for the period 2003-2007 using observed rates until the end of 1987. Two short term estimates were calculated for the data; one for the period 1993-1997 for the observed data until the end of 1987, and the other for the period 2003-2007 for observed data until the end of 1997. The comparison of the two short-term predictions allows for comparisons to be made on the consistency of the estimation approaches over time.

4.5.8. Potential for Dampening

Another proposition that has been made as part of previous comparisons of the available methods is the tempering (or dampening) of the projected functions for longer-term projections [Møller et al., 2003]. Although not directly compared in this chapter, there is the opportunity to extend each of the methods proposed to incorporate some form of dampening. The aim of this chapter is to establish the effect of treating time continuously compared to partitioning the time-scales into five-year intervals, and to assess the effect of using splines to incorporate more recent time-trends as opposed to projecting the full drift. Each of the methods will equally be improved by dampening, and therefore it would not add anything to the comparisons made here.

4.6. Application

Table 4.1 shows the comparison of the different methods in terms of the average absolute value of the total relative difference between the observed and expected cases for each of the cancer sites for the period from 1993 until the end of 1997. Methods using the same link function for the three overall approaches ("Spline Restriction", "Projecting Drift", and "Factor Drift") should be directly compared. Also, comparisons within each of the methods for the difference made by the selection of the link function is appropriate. The "Factor (Drift)" method (Mean of 12.02 for "Log", and 11.78 for "Power"), which is the usual method for making the projection,

performs worse on average than each of the other methods ("Spline Restriction"; Mean of 8.12 for "Log", and 6.68 for "Power", and "Projecting Drift"; Mean of 10.44 for "Log", and 9.73 for "Power") for both the "Log" and "Power" models.

In the specific cases for each cancer site, split also by gender, the results in Table 4.1 show that the Spline Restriction method gives better predictions than the Factor Drift method for both examples of the link function in the majority of cases. For the log link function, 4 of the 7 estimates for the spline restriction method give an improvement over the standard factor model. For the power link function, the spline restriction approach is better in all 7 cases. Improvements can also be seen when comparing the factor model to the model that projects the full drift in a continuous fashion.

Link Function	Log			Power (5)		
Cancer Site	Spline Restriction	Projecting Drift	Factor Drift	Spline Restriction	Projecting Drift	Factor Drift
Breast (Females)	5.66	2.36	8.91	9.93	8.91	14.81
Colon (Males)	12.20	4.97	6.25	5.64	3.50	7.30
Colon (Females)	11.26	13.70	10.18	4.05	5.59	4.26
Lung (Males)	6.46	14.60	19.15	12.29	21.26	25.97
Lung (Females)	9.56	7.48	6.70	4.62	2.91	4.74
Pancreas (Males)	6.23	14.74	18.14	6.73	14.80	15.91
Pancreas (Females)	5.44	15.27	14.82	3.51	11.17	9.45
Mean	8.12	10.44	12.02	6.68	9.73	11.78

TABLE 4.1. Observed data until the end of 1987; 10 year prediction for the period 1993-1997. The figures give the average yearly absolute relative difference (%) between the observed and predicted number of cases for all ages combined.

Table 4.2 contains the results for the longer-term predictions for the period from 2003 to 2007. It was expected that for longer-term predictions the exponential growth that is introduced by the logarithmic link function may lead to over-estimates of the projected incidence, and that the suggested alternative of the power link function (with a power of 5) may well yield better predictions. However, it is apparent from the results in Table 4.2 that this is not the case for all of the cancer sites. The estimates for lung cancer for males is a particular example of the power link giving substantially poorer estimates than the log link function. Although, in general the power function does give a better fit, it is not necessarily better in every scenario.

Figure 4.3 gives the graphical representation of the results for the pancreatic cancer data for females. The data plotted in the graph are the predicted total number of cases for all ages combined from each of the approaches compared to the total number of observed cases for all

Link Function		Log			Power (5)	
Cancer Site	Spline Restriction	Projecting Drift	Factor Drift	Spline Restriction	Projecting Drift	Factor Drift
Breast (Females)	5.04	2.20	10.40	15.50	17.24	23.49
Colon (Males)	26.34	6.80	7.56	12.48	5.57	8.39
Colon (Females)	29.44	38.74	32.77	11.50	16.81	10.77
Lung (Males)	10.41	28.07	35.41	26.50	53.74	59.75
Lung (Females)	6.28	5.58	5.42	4.44	9.26	12.99
Pancreas (Males)	12.11	6.70	11.15	7.14	8.31	8.30
Pancreas (Females)	4.81	19.65	17.40	5.37	9.75	6.97
Mean	13.49	15.39	17.16	11.85	17.24	18.66

TABLE 4.2. Observed data until the end of 1987; 20 year prediction for the period 2003-2007. The figures give the average yearly absolute relative difference (%) between the observed and predicted number of cases for all ages combined

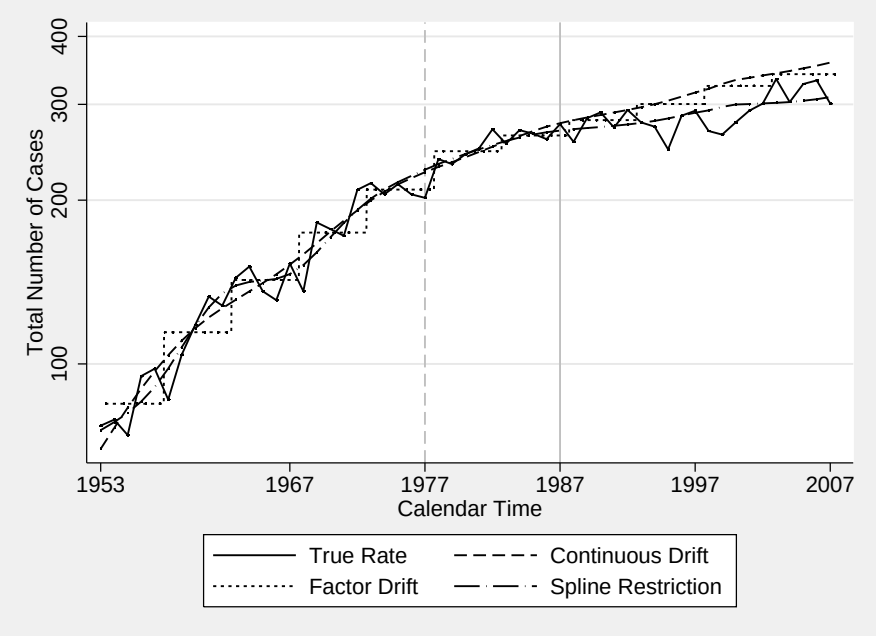

FIGURE 4.3. Projections from 1987 for female pancreatic cancer patients for the total number of cases for all ages. GLM fitted with a power link function.

ages combined. The figure shows the fitted curves for the 3 methods for the power link function, as well as the observed values of the total number of cases. The 10 year window used for the spline restriction shows a gradient that is lesser than that observed over the longer observation window. Therfore, the spline restriction approach gives better projections after 1987 in this example. Using the "recent" trend gives a better method of estimating the future trend than using the longer-term "drift" to make the projection.

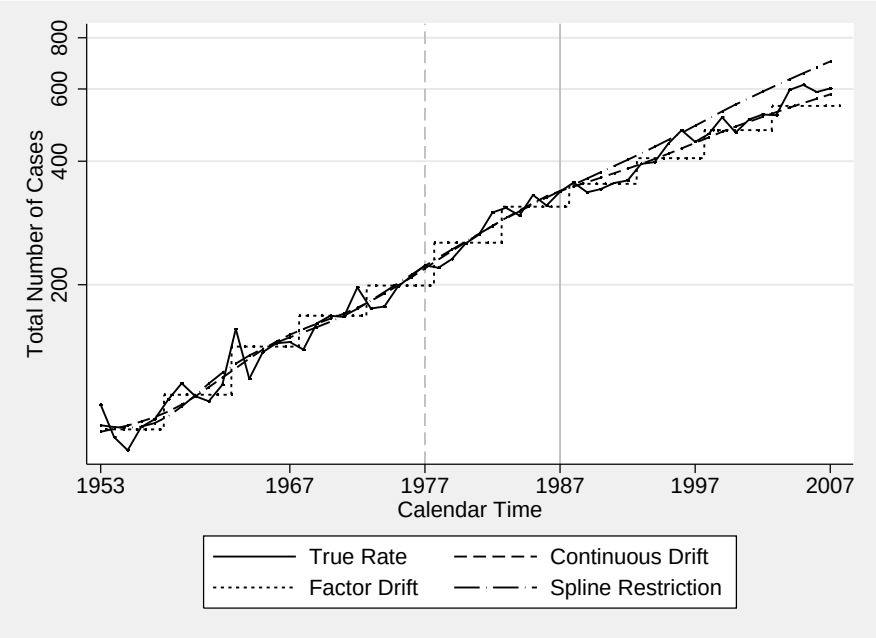

FIGURE 4.4. Projections from 1987 for male colon cancer patients for the total number of cases for all ages. GLM fitted with a power link function.

In contrast, Figure 4.4 gives the graphical representation of the results for the colon cancer data for males. Again, this figure compares the methods that use the power link function. The drift projection methods that use the drift over the entire observation period ("Drift Projection" and "Factor Drift") clearly perform well in this case, which can be seen from the values given in the relevant tables (Tables 4.1 and 4.2). In the case for males, the ten year window used for the spline restriction method shows a gradient that is larger than observed over the entire observation window as a whole. However, this gradient does not continue past 1987 when the projections are made. This is an example of when taking the "recent" trend does not give a better projection than the overall "longer" trend.

Figure 4.5 gives the graphical representation of the results for the male lung cancer patients using a logarithmic link function. This is one of the examples where the log link function outperforms that of the power link. This provides a perfect example to illustrate the dangers of using the longer-term drift in order to project into the future. There is a clear change in the pattern of the lung cancer cases over time for the males with a substantial decrease in the incidence from the 1970s onwards; this is highly likely to be due to a change in smoking habits

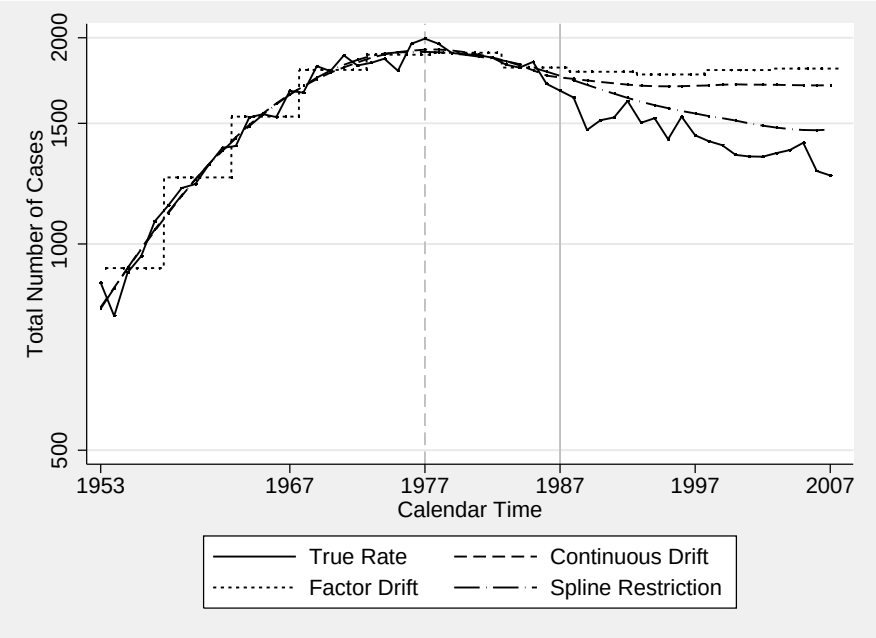

FIGURE 4.5. Projections from 1987 for male lung cancer patients for the total number of cases for all ages. GLM fitted with a log link function.

for Finnish males. This leads to the method using the spline restriction, which is dominated by the change in the last 10 years, to outperform the "full drift" approach.

Link Function	Log			Power (5)		
Cancer Site	Spline Restriction	Projecting Drift	Factor Drift	Spline Restriction	Projecting Drift	Factor Drift
Breast (Females)	8.75	2.60	2.41	4.61	4.12	8.19
Colon (Males)	5.25	8.43	8.25	4.68	4.39	4.69
Colon (Females)	8.95	15.77	14.56	5.58	8.40	6.55
Lung (Males)	9.83	3.20	5.24	3.37	15.73	19.50
Lung (Females)	10.77	3.83	6.78	11.79	7.57	11.01
Pancreas (Males)	16.72	10.58	6.44	13.98	9.46	6.21
Pancreas (Females)	5.44	5.16	4.49	5.39	5.32	5.05
Mean	9.39	7.08	6.88	7.06	7.86	8.74

TABLE 4.3. Observed data until the end of 1997; 10 year prediction for the period 2003-2007. The figures give the average yearly absolute relative difference (%) between the observed and predicted number of cases for all ages combined

Table 4.3 contains the projections for data observed until the end of 1997, and projected until the end of the available data in 2007. This table is included to assess the consistency of the three methods of estimation. It is clear from the results contained in Table 4.3 that there is a lack of consistency across time-points for the "best" method of estimation for any given

cancer site. The methods are likely to give similar estimates if the recent trend in incidence rates is similar to that of the overall trend. This seems to be largely the case for this later time period.

4.6.1. Sensitivity Analyses

Two separate sensitivity analyses were carried out. Firstly, the sensitivity to the number of knots when projecting is compared. The main emphasis of the last chapter was to assess the impact of knot placement when dealing with observed data only. Secondly, a sensitivity analysis is conducted to investigate the choice for the placement of the final boundary knot.

4.6.1.1. *Number of Knots*

A common criticism of the use of spline functions is the arbitrary nature of selecting the number and position of the knots. The results of a sensitivity analysis carried out in this setting are contained in Table 4.4. For the sensitivity analysis, the degrees of freedom for the age and cohort spline terms are kept constant whilst varying the degrees of freedom for the period term. The analysis for Table 4.1 was performed for each of the new models with the varying set of knots for period (degrees of freedom of 3 (Model A), 5 (Model B), 6 (Model C), and 8 (Model D)) and with 8 degrees of freedom for the age and cohort terms. The degrees of freedom used to carry out the original analysis were equivalent to Model B.

Table 4.4 shows that for the majority of cancer sites varying the degrees of freedom for period makes little difference to the estimated mean value of the percentage relative difference. However, for "Colon (Males)" and "Lung (Females)" quite substantial differences occur. In the case of colon cancer for males, the models with an increasing number of knots for period seem to give poorer estimates than the simplest model (Model A). Figure 4.4 shows that there seems to be quite a simple relationship between the total number of cases, and calendar time. From this, it is possible to conclude that the other models are overfitting the effect of period. This is confirmed by looking at the AIC and BIC, which show that the model with 3 degrees of freedom for period is the best fitting model for the observed data. For female lung cancer, the opposite seems true. It seems as though the simpler models are underfitting the effect of period and that increasing the degrees of freedom leads to better projections. However, this is not confirmed when comparing the values of the AIC/BIC. A "better" fitting model to the observed data does not necessarily lead to better projections. In real analyses when projections

	Link	-	-	-	Log	Power	Log	Power
		Age	Period	Cohort	Spline	Spline	Projecting	Projecting
Model	**Cancer Site**	(df)	(df)	(df)	Restriction	Restriction	Drift	Drift
(A)	Breast (Females)	8	3	8	5.30	10.29	6.39	12.63
(B)		8	5	8	5.66	9.93	2.36	8.91
(C)		8	6	8	5.96	10.08	1.97	8.04
(D)		8	8	8	4.69	8.69	1.87	7.36
(A)	Colon (Males)	8	3	8	8.96	3.85	5.82	3.41
(B)		8	5	8	12.20	5.64	4.97	3.50
(C)		8	6	8	13.73	6.47	4.29	4.04
(D)		8	8	8	17.81	10.20	3.71	5.70
(A)	Colon (Females)	8	3	8	16.81	8.09	14.98	6.23
(B)		8	5	8	11.26	4.05	13.70	5.59
(C)		8	6	8	12.68	5.64	13.61	5.28
(D)		8	8	8	12.18	4.90	11.55	3.73
(A)	Lung (Males)	8	3	8	8.18	14.87	15.18	22.53
(B)		8	5	8	6.46	12.29	14.60	21.26
(C)		8	6	8	7.16	12.71	14.09	20.80
(D)		8	8	8	7.76	12.86	16.00	23.21
(A)	Lung (Females)	8	3	8	14.88	8.71	9.18	2.63
(B)		8	5	8	9.56	4.62	7.48	2.91
(C)		8	6	8	7.24	3.56	9.84	2.72
(D)		8	8	8	4.38	2.79	12.12	3.87
(A)	Pancreas (Males)	8	3	8	6.69	7.12	11.56	10.69
(B)		8	5	8	6.23	6.73	14.74	14.80
(C)		8	6	8	8.10	8.75	14.98	14.82
(D)		8	8	8	5.48	7.26	17.10	17.00
(A)	Pancreas (Females)	8	3	8	12.92	10.26	15.06	10.29
(B)		8	5	8	5.44	3.51	15.27	11.17
(C)		8	6	8	6.28	3.76	14.21	11.07
(D)		8	8	8	10.64	7.55	10.32	7.54

TABLE 4.4. Models (A)-(D) relate to the different choice of degrees of freedom (df) for Period. The values given are equivalent to the values in Table 4.1. They relate to average yearly absolute relative difference (%) between the observed and predicted number of cases for all ages combined for the 10 year projections from 1987. The degrees of freedom for Model (B) were the ones used in the actual analyses. The knots were equally spaced at the centiles of the relevant variables.

are actually made into the future rather than into a period where the number of cases is known, it will not be possible to compare the degrees of freedom in this way. It is therefore essential to use care and consider whether the projections made in any given scenario appear sensible, and that they align with any external knowledge about the disease of interest.

4.6.1.2. *Placement of the boundary knot*

The new approach using the restriction of the cubic splines has been proposed with the further condition that the boundary knots for the period and cohort terms are brought within the range of the data. Moving the boundary knot further into the known data will reduce the recentness of

the data included for the projection. Moving the boundary knot to the extreme of the observed data may well lead to more unstable projections. For the analyses conducted in this chapter, the boundary knot was moved in 10 years. However, it is possible to perform a sensitivity analysis to see how the projections will vary depending on where the boundary knot is placed.

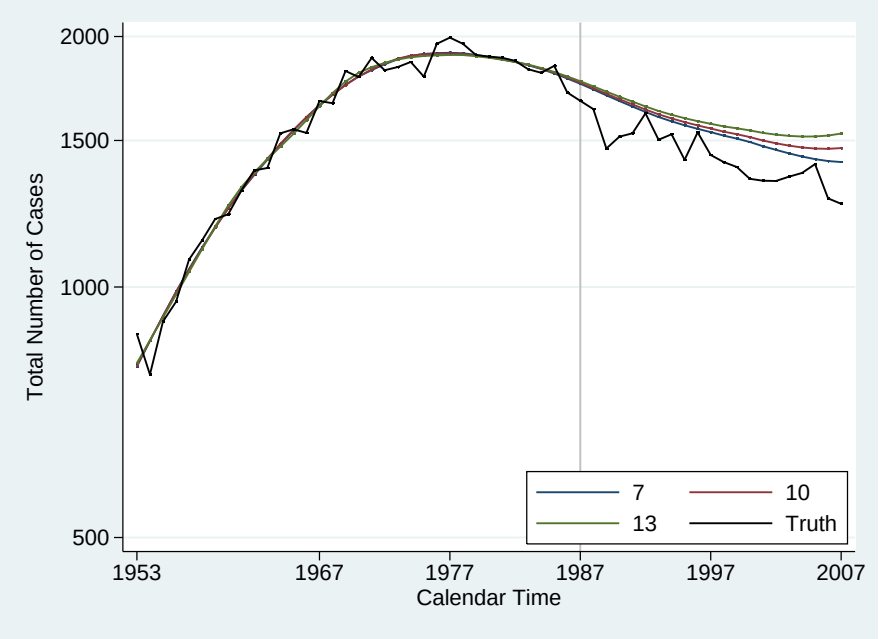

FIGURE 4.6. Projections from 1987 for male lung cancer patients for the total number of cases for all ages. The different lines correspond to moving the boundary knot to different points within the range of the data.

Figure 4.6 shows the plot for lung cancer where the boundary knot has been moved 3 years either side of the 10 year value that was used in the analyses. It is clear that there is not too much sensitivity to the value that is selected over a small range of years. The projections only seem to diverge towards the end of the projection period, where there is greater uncertainty about the continuation of the linear trend.

Figure 4.7 shows the plot for lung cancer where the boundary knot has been placed within the data over a larger range of values (from 0 to 15 years). Letting the boundary knot be closer to the end of the data for lung cancer results in the projection being closer to what occurred because of the fact that the decrease in incidence was still observed beyond 1987. The lines produce a fan shape for the projected lines from the point of 1987. If the "fan" of lines is very tight and there is not much variability in the projected estimates, this would suggest that the

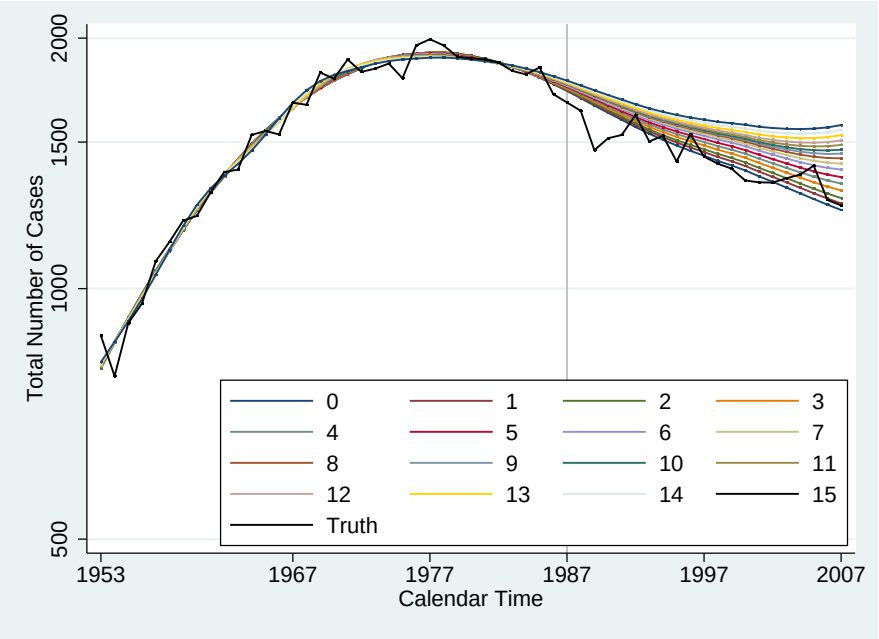

FIGURE 4.7. Projections from 1987 for male lung cancer patients for the total number of cases for all ages. The different lines correspond to moving the boundary knot to different points within the range of the data.

full drift approach is likely to give similar estimates to the spline restriction approach. However, if the fan spreads out such as in Figure 4.7, this suggests that there is not a single linear trend over the entire range of observed data, and then a choice is required as to which data to include when making the projections.

Figure 4.8 shows the plot for colon cancer where the boundary knot has been moved 3 years either side of the 10 year value that was used in the analyses. Again, it is clear that there is not too much sensitivity to the placement of the boundary knot over a small range of values. Moving the boundary knot over a small range of values within the range of the data will not result in wildly different projection estimates. In each case, the linear projection made is based on the data after the placement of the boundary knot. Provided that the boundary knot is not placed at the very edge of the available data, each of the projections made will be dicatated by the linearity in the latter part of the available data. If the boundary knot is placed too far into the range of the data, this linearity constraint may be unrealistic and lead to poor projections. However, this is not the case for colon cancer for males in Finland.

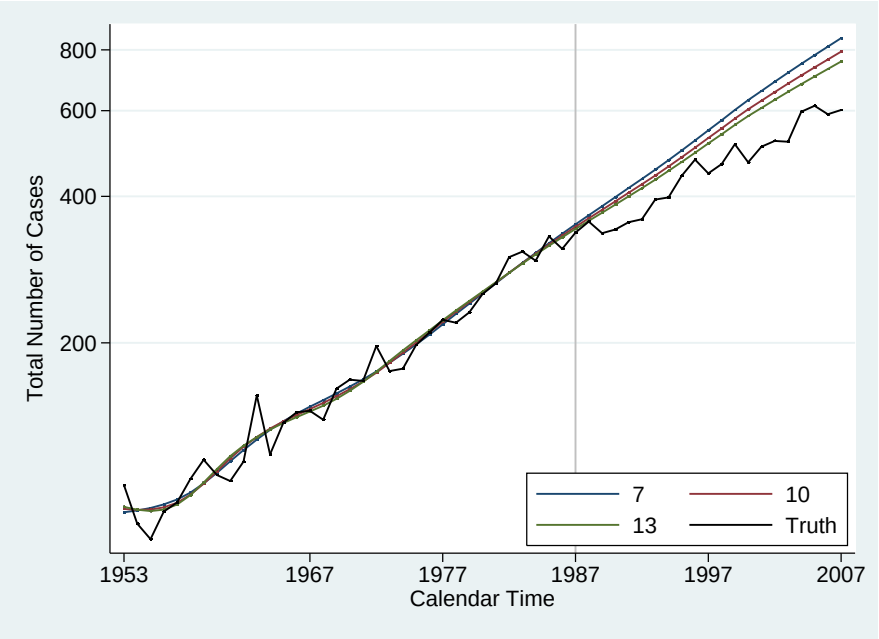

FIGURE 4.8. Projections from 1987 for male colon cancer patients for the total number of cases for all ages. The different lines correspond to moving the boundary knot to different points within the range of the data.

Figure 4.9 shows the sensitivity analysis plot over a larger range of values for the male colon cancer data. Steps of 5 year intervals for the movement of the boundary knot into the range of the data have been used in this instance as the better fitting projections result from using an extended linear period for the cohort and period term, up to 30 years within the range of the data. Using such a long period of linear constraint is similar in principle to projecting the linear drift into the future; which is why the spline drift approach performed well for the male colon cancer data. These plots that produce a fan of potential scenarios can be useful for assessing the changes over time in terms of incidence trends, as well as for expressing the uncertainty underlying the projections.

4.7. Discussion

The methods that treat the time-scales in a more continuous way give "better" projections than the factor model in the majority of scenarios. If yearly data are available then it is wasteful to use the factor models in five-yearly splits of the time-scales. Of the two spline methods, the spline restriction approach that uses a more recent trend to make the projections often performs

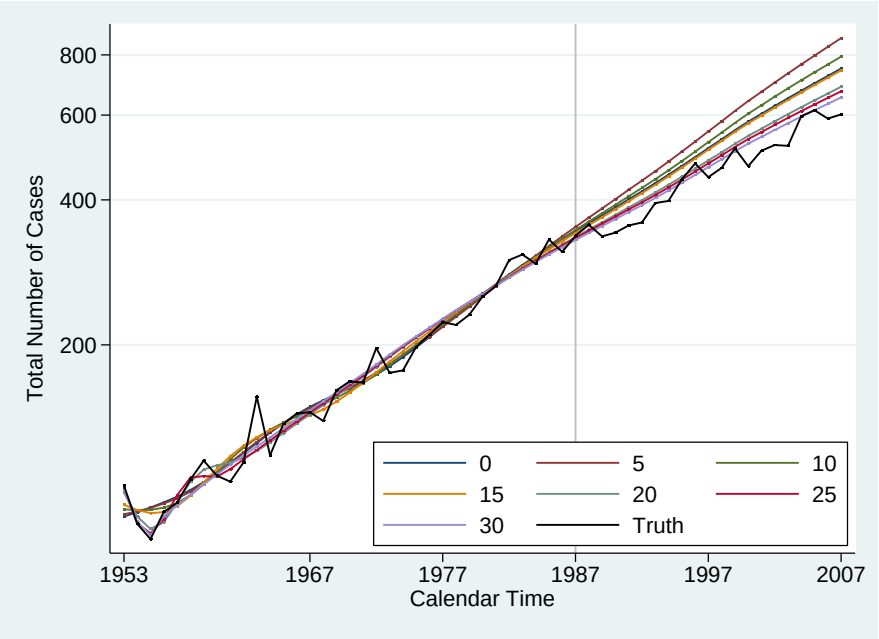

FIGURE 4.9. Projections from 1987 for male colon cancer patients for the total number of cases for all ages. The different lines correspond to moving the boundary knot to different points within the range of the data.

better than the full drift approach especially in scenarios where there has been a change in the shape of the incidence curve over the entire period (see next chapter for further examples).

The measure used in these analyses is a more appropriate measure to use if interest lies in effectively measuring the predictive ability of the methods on a yearly basis. It seems as though, in practice, that it would be of greater interest to have "good" projections in yearly intervals rather than in five-year time windows. If interest lies in the average effect over a five-year window, it is still possible to obtain this from the spline techniques, and it has been shown that "better" estimates over a five-year period are obtained from using yearly data. The corresponding values (Table 4.1 - Power link function) for the mean values over all sites using the average effect over a five-year window are 9.22 for the spline drift approach and 10.66 for the factor drift approach. There is still a benefit for adopting a smoother approach even if an average value over a period is required.

The extrapolation of the incidence rates for each of the methods are based on the assumption that the trends in the past will continue into the future. The error that this introduces can be seen by comparing the estimated incidence rate curves with the true, observed incidence

rates. Using these methods outside of a retrospective setting would lead to this error being an unknown quantity and caution should be taken when making any extrapolation. However, projections are often important for the planning of services and these methods are useful if appropriate caution is taken. The "one method fits all" approach to projecting incidence rates is not an approach that should be advocated. This is a point that has not be made clear when large-scale comparisons of the available methods have been carried out [Møller et al., 2003]. Each cancer site, and even each time-point within a cancer site, may require a different approach for making a sensible projection. However, without the benefit of hindsight, it can be difficult to select the most appropriate method of projection. Care must be taken to look at the shape of the incidence rate up to the point of projection and to give particular attention to the shape of the curve towards the end of the observation period.

For the projections made in this analysis, true population estimates were used because the analysis was done on historical data. For future predictions of the number of cases, further errors will be introduced by the inaccuracies in the forecasting of population data. The methods for forecasting population data are subject to some of the same issues of the projection of the incidence rates. It is necessary to make assumptions about the birth and death rates for the populations as well as assumptions about the level of immigration and emigration. However, for the majority of countries, projections of population figures are fairly accurate.

Further work is necessary for providing a relevant measure of uncertainty for the predictions that are made for each of these methods. There is a great deal of uncertainty in the estimates, and this may not be fully appreciated when these projected rates are presented. It has been suggested that looking at a range of scenarios can give an idea of the uncertainty of the estimates. It is often proposed that assuming that the rates will stay the same as the last observation point is a suitable lower/upper bound for the projections.

The projection method that uses the restriction of the cubic splines can be sensitive to the placement of the boundary knot. Further investigation on the length of time that should be linear within the range of the data is included in the second part of the sensitivity analysis section. If the boundary knot is placed at the very end of the observed data this can lead to local trends having too much of an influence on the projections. The 10 years that is suggested

here is adequate for averaging out any local deviations. However, using too large a value for this length may well detract from the "recentness" of trends that this method can suitably capture.

Using restricted cubic splines for either the drift projection approach or the approach that uses the linear restriction requires the specification and placement of the knots for the splines. This is an issue that has been given a lot of consideration for various applications using splines. Most of these assessments conclude that provided a sufficient number of knots are used then the results are not overly sensitive to the number and placement of the knots. The sensitivity analysis carried out in the first part of Section 4.6.1 highlights that there is some sensitivity to the placement of the knots, particularly for the newly proposed method. Equally spacing the knots throughout the interval appeared to be the most appropriate method for the placement. Specifying too many knots may well result in overfitting of the incidence rates which could lead to spurious projections for the linear restriction method if care is not taken with the boundary knot placement.

The method using the restriction of the cubic splines is similar in principle to using a prediction based on a more recent estimate of the drift. This is a method that was recommended in the empirical comparison carried out by B. Møller *et al.* [2003]. In this chapter, this is put in a setting that treats the effects of age, period and cohort continuously. The method proposed here also allows for a simple comparison between the "recentness" to use for the projection by simply moving the boundary knots for the restricted cubic splines. The suggestions of using a power link function, and halving the drift after 10 years can easily be applied to the outlined approach.

Further validation of the projection approach using restricted cubic splines is required. It is intended that the new approach will be applied to UK cancer data and compared to the results obtained by H. Møller *et al.* [2007]. This will allow the method to be compared using a larger population size than that available with the Finnish data. The method is outlined in a paper that is currently undergoing peer-review [Rutherford et al., 2011b].

CHAPTER 5

Dangers of Incidence Projections

5.1. Chapter Outline

In this chapter, a number of examples are given of when incidence projection methods perform poorly. The Finnish cancer registry data is used to highlight a number of sites where at least one of the methods compared in the previous chapter gives poor projection estimates. Throughout the chapter, discussion is given to the care that is required when using these approaches. Through careful thought, data exploration and, in some cases, external information it is possible to lessen the chance that an inappropriate method will be applied; leading to improved projections.

5.2. Introduction

Making projections of incidence and other measures of disease burden are of great importance for health planning authorities so that the appropriate finance and services are provided [Theisen, 2003]. There have been numerous methods suggested to carry out the projections based on the data available, and a selection of the methods have been evaluated in the previous chapter. In this chapter, the dangers of making projections using these techniques are highlighted; in some cases, the projections can be severely wrong. This will highlight the extreme caution that must be taken when making projections. Although it is not possible to assess the assumptions that are made until the data eventually becomes available, it is possible to use external information and experience as a guide. In this chapter, a number of examples are given using Finnish Cancer Registry data that show how each of the methods discussed in the previous chapter can "go wrong" and discussion is given as to how this can be avoided, or at least how to lessen the danger of it occuring.

5.3. Methods

The methods described in the previous chapter for making projections with age-period-cohort models are applied to various cancer sites at purposefully selected time-points to show examples of "worst-case" scenarios for each of the techniques. The methods that use the drift projection will perform poorly in some scenarios where the spline restriction method will perform well, and vice versa. In some cases, both methods will perform poorly; this could be due to, for example, the introduction of a new screening program just after the last observed data. The introduction of a screening program could initially inflate the incidence estimates for the screened population. This has seen to be the case for breast cancer and mammography screening [Quinn and Allen, 1995] and has also been investigated for prostate cancer and PSA screening [Pashayan et al., 2006]. Methods to account for the effect of mammographic screening on breast cancer incidence projections in Finland have been proposed [Seppänen et al., 2006]. The projection models are made under a variety of scenarios for potential future screening practices for Finland.

In order to highlight these examples of poor projections, a retrospective analysis was used (similar to the one applied in the previous chapter to compare the methods). Models are fitted to known data for a given time interval and projected forward using the various techniques. The true rates are then compared to the projections graphically. Graphical representations of the total number of cases estimated for each method are given for every example. Due to the fact that a retrospective analysis has been applied, it is also possible to plot the true number of cases. These figures are similar to those seen in the previous chapter when evaluating the best performing methods. Further to these figures, it is also possible to produce plots that show the the underlying assumptions made when using the different projection methods by returning to the more traditional APC plots that were introduced in Chapter 2. These figures are somewhat harder to interpret. However, they give a true reflection of the assumptions made by each of the methods, and have the potential to highlight the reasons for any differences between the techniques. For each of the spline models that are compared in the results section, 5 degrees of freedom are used for the period term, whereas 8 degrees of freedom will be used for the age and cohort terms.

5.4. Data

The cancer sites that have been selected have been somewhat "cherry-picked" in order to highlight how poorly the projection methods can perform in certain scenarios. However, there have been numerous examples whereby a "one method fits all" approach has been applied across cancer sites in national and international studies [Engeland et al., 1993; Agha et al., 2006; Smith et al., 2009]. This highlights that even when a method could have been predicted to perform badly, it has still been applied. In some circumstances, measures have been taken to account for the fact that one method may perform badly in some scenarios. For example, the trend over time has been tested for linearity to see if it is sensible to apply the approaches that project the overall drift [Møller et al., 2007]. The following examples highlight that even with precautionary measures, there are still circumstances where projections can go wrong. This illustrates the level of uncertainty that surrounds the projected estimates and highlights the care that must be taken when making projections.

As in the previous chapter, up to 20-year projections are given in each example. Given that, at the time of writing, the Finnish Cancer Registry data is only available until the end of 2007, projections are made from 1987. The analyses of projecting incidence are done separately for each sex. The selected cancer sites to show the dangers of incidence projections are lung cancer for males, non-Hodgkin's lymphoma for females, cancer of the rectum for females, testicular cancer for males, and finally Hodgkin's lymphoma for males. In the previous chapter, the most common cancer sites were chosen to illustrate the methods on the basis that Finland is a relatively small country. In this chapter, the comparisons of the methods are made for both common cancer sites (lung, non-Hodgkin's lymphoma and rectum), and less common sites (Hodgkin's lymphoma and testis).

5.5. Results

Figure 5.1 shows the two spline methods of projection for the total number of cases of lung cancer for males. The models used to produce this graph used the standard log link function for the Poisson GLM. It is evident that in this instance the projection method using the restriction of the cubic splines gives a better estimate of the true rate after 1987. It is also evident from this graph why that is the case. This is a prime example of when using the long-term drift is clearly

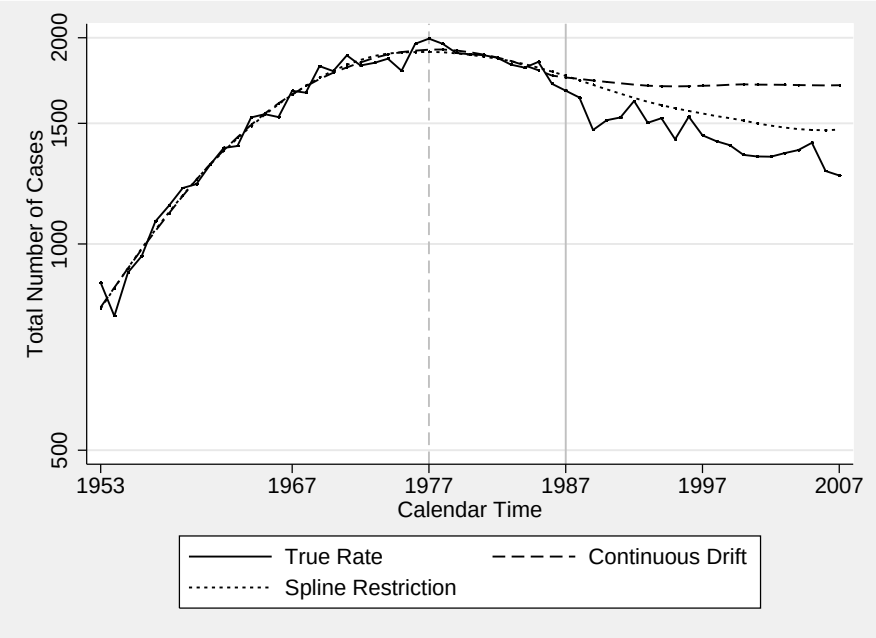

FIGURE 5.1. Projections from the two approaches from 1987. Lung cancer for males. A logarithmic link function was used for the projection models.

inappropriate. Due to the almost parabolic shape of the fitted curve up until 1987, the linear drift term is effectively zero. However, due to the non-linear shape of the curve, the average linear trend is not representative of the gradient of the curve for large parts of the observed data. There is a clearly non-linear shape to the curve, and it seems logical based on the data prior to 1987 to assume that this will continue. This is effectively what is achieved by using the spline restriction method. This downturn in the lung cancer cases for males is probably a reflection of the smoking trends for Finnish males over time. Therefore, it is likely that at some point in the future the curve for the total number of cases will reach a plateau. If this were to have occurred from exactly 1987 then the full drift approach would have given better projections; however, this would have been more by lucky coincidence than accurate capturing of the shape of the available data.

Figure 5.2 shows the age, period and cohort effects separately whilst highlighting the underlying assumptions made by each method. In this illustration, the drift term has been attributed to the period effect on the basis that this feels the more natural timescale when making the projections. The age effects are given on a rate scale, whereas the period and cohort terms are

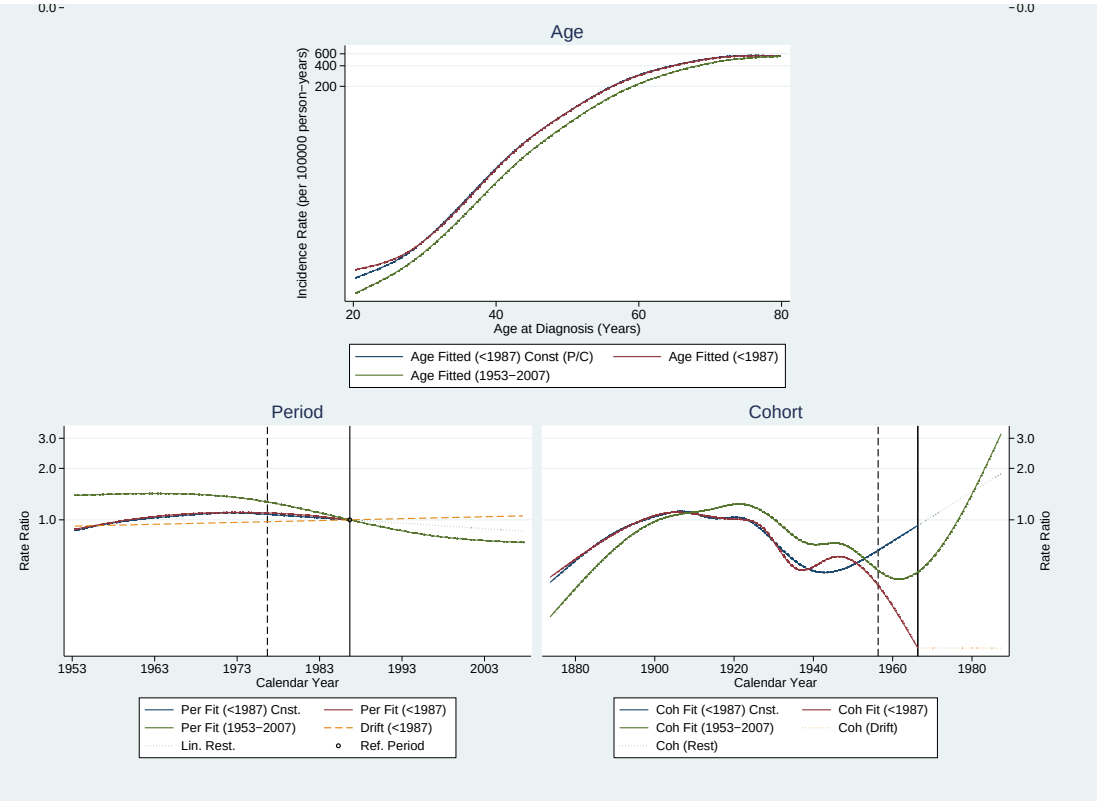

FIGURE 5.2. Age-period-cohort graph illustrating the two approaches. Lung cancer for males.

given on a rate ratio scale. Within each of the age, period and cohort graphs, there are three fitted functions. The solid line at Calendar Year= 1987, and Calendar Year= 1967 for period and cohort respectively highlights the point at which the projections are made for each component. The 20 year difference between these two points is equivalent to the age of the youngest patient contained in the analysis. The age effect is assumed to be the same in the future as it has been up until the point at projection. The fitted functions given in blue relate to the model that uses the spline restriction method to project the incidence into the future. For these functions, the boundary knots have been brought into the range of the data by 10 years so that the restriction of the cubic splines is applied from an earlier point in time. This leads to a linear projection for both the cohort and period terms, with the gradient of this projection being defined by the latter part of the relevant cohort and period data. An equivalent projection is made if the drift is allocated to the cohort term instead of the period term. The fitted functions given in red (projections in orange) relate to the model that uses the drift to project the incidence into the

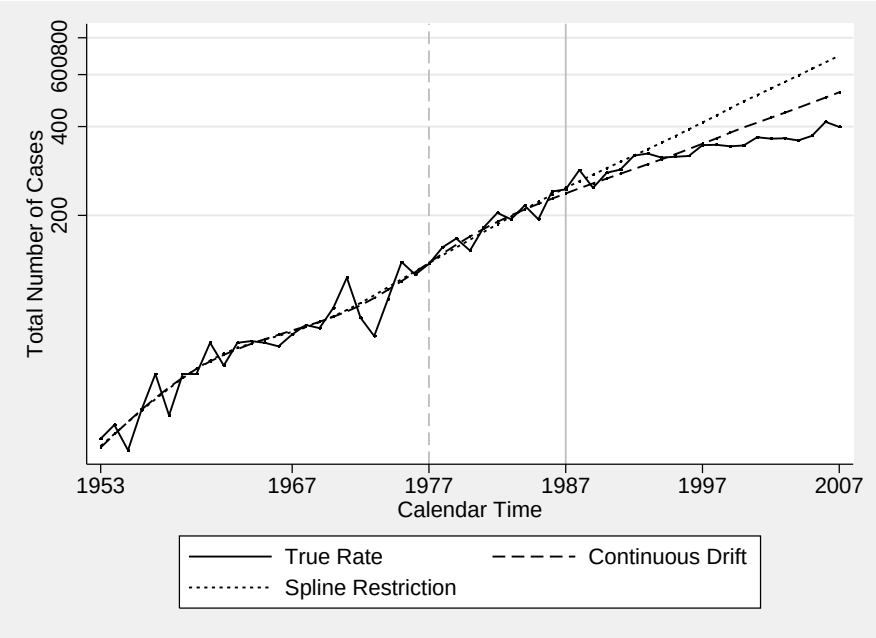

FIGURE 5.3. Projections from the two approaches from 1987. Non-Hodgkin's lymphoma for females. A logarithmic link function was used for the projection models.

future. In this case, an age-period-cohort model using restricted cubic splines is applied to the data prior to 1987, and the fitted function for age, period and cohort are displayed graphically in Figure 5.2. Beyond 1987, the drift that is estimated from the APC model is used to project the period function in the future. The straight dashed line shows the line with the estimated drift as the gradient and an appropriate constant term so that the straight line intersects the fitted curve at 1987. For the future cohorts, the fitted function is assumed to be constant at the level of the last fitted cohort. There are distinct similarities between the shape of the period curve of Figure 5.2 and the curve illustrating the total number of cases for male lung cancer over calendar time given in Figure 5.1. The fitted function for the APC model applied to the entire range of the data are given in green. This is purely for comparative purposes, and uses the data beyond the projection point so that a comparison of the fitted functions can be made. However, it is difficult to directly compare the curves from this model purely because of the range of the data for which the constraint applies. For the model that is fitted over the entire range of the data (1953-2007) the period and cohort terms are constrained over a longer range than for the other two fitted functions, which are constrained over the period 1953-1987. This

can explain some of the differences in the cohort terms in Figure 5.2. The large differences at the end of the cohort term do not have a large influence on the total number of cases. This is because of the fact that those born in the more recent cohorts are the youngest patients who have a much lower rate of incidence.

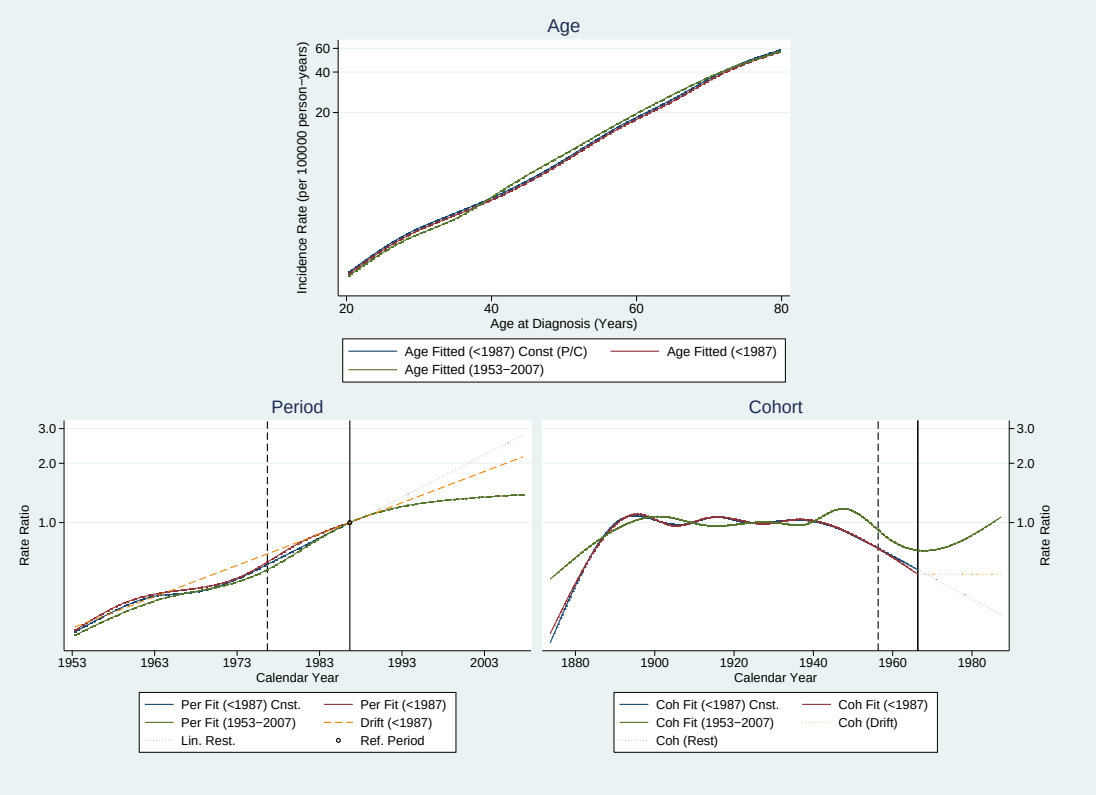

FIGURE 5.4. Age-period-cohort graph illustrating the two approaches. Non-Hodgkin's lymphoma for females.

Figure 5.3 again shows one of the projection methods performing considerably better than the other. In this instance, the method using the full drift gives better projections in the long-term, whereas in the short-term (up to 6 years post-prediction) the spline restriction method performs well. The figure suggests that there is an upturn in the total number of cases of non-Hodgkin's lymphoma for females from around 1977. This change in gradient appears to last up until around 1992-3, before there is a return to the long-term average linear trend. Without using hindsight, it is difficult to imagine anyone being able to predict this based on information available at the point of projection (that is, in 1987). Again, this shows the dangers of making projections based on simple assumptions that may or may not hold.

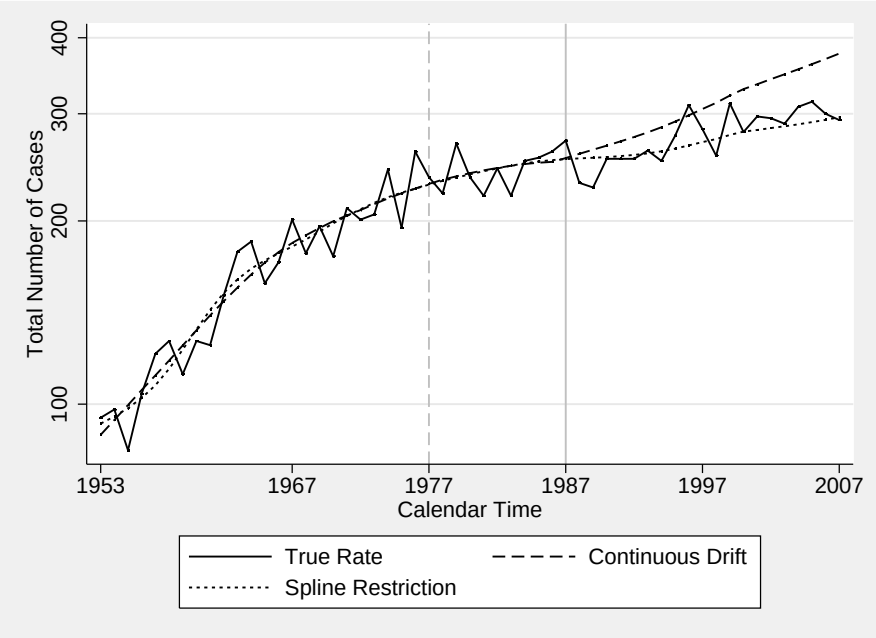

FIGURE 5.5. Projections from the two approaches from 1987. Cancer of the rectum for females. A logarithmic link function was used for the projection models.

Figure 5.4 confirms that the drift approach is closer to the "true" function for both period and cohort. Although these APC graphs are difficult to interpret, they do give a clear indication of what the assumptions of the methods are for each of the approaches. The similarities between the cohort and age curves for all three of the fitted functions, results in the curves for period following a similar pattern to those observed in Figure 5.3. The differences that are observed for the youngest cohorts do not have a great influence on the total number of cases due to the fact that these patients have a low incidence rate.

The next example is for rectum cancer in females; the results are contained in Figure 5.5. Again, we see a non-linear shape for the graph. The spline restriction approach outperforms the full drift approach because the trend over the last 10 years is lower than that over the 34 year observation period, and this new trend is the one that continues into the future.

Figure 5.6 shows the APC graph for female rectum cancer. The cohort effect after 1966 does not have a great influence on the total overall incidence rate because of the fact that rectum cancer has less incident cases for the younger ages. This results in the projection being made from the period curves being dominant in dictating which of the methods performs the

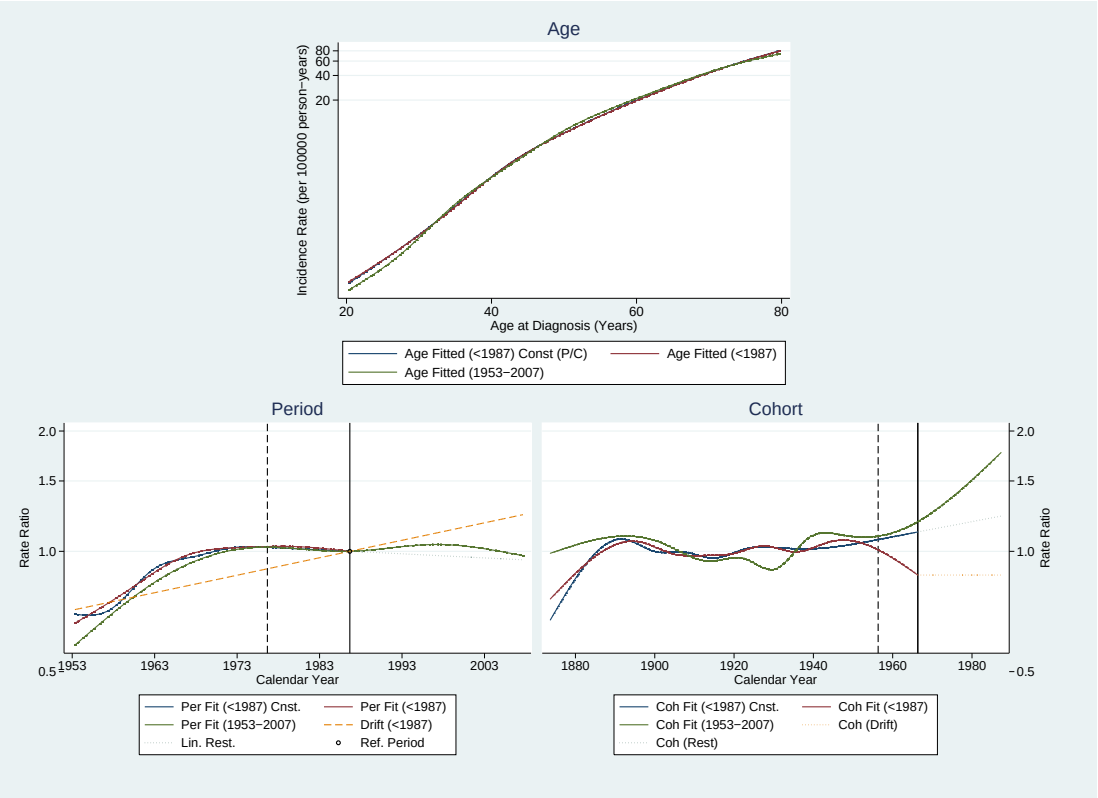

FIGURE 5.6. Age-period-cohort graph illustrating the two approaches. Cancer of the rectum for females.

best. It can be seen that the restriction approach seems to fit better to the true data, and that the fact that the cohort effect is overestimated by that approach means that the overall projected rate is even closer to the truth.

Figure 5.7 shows the results for testicular cancer for males. In this example it appears as though the relatively small number of cases leads to the starting value for the lines for projections at the end of 1987 to differ between the two methods. This results in the drift approach overestimating the total number of cases for the prediction window. The fitted splines would be less likely to "pick up" local deviations if fewer knots were used for the splines. Also, the local deviations would be less likely for both a more common cancer site, and also for the same cancer site in a country with a larger population.

Figure 5.8 seems to suggest that the continuous drift approach should give better estimates than the spline restriction approach, which is contrary to what is seen in Figure 5.7. However, the difference in the age curves, which are the age-specific rates for the reference period, is the

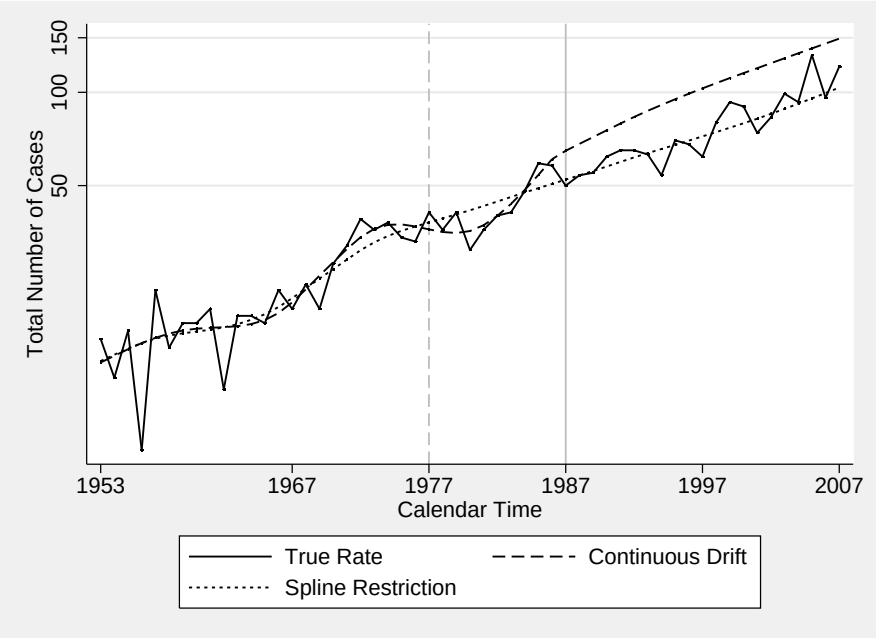

FIGURE 5.7. Projections from the two approaches from 1987. Testicular cancer for males. A logarithmic link function was used for the projection models.

reason behind the apparent disagreement. This difference is the difference that is expressed between the two starting points for the projection at the point of 1987 in Figure 5.7. In the previous APC graphs given in the examples, the age curves have been broadly similar for the two projection methods. This is because, usually, the period effect at 1987 is usually estimated to be the same from both approaches. In the case of testicular cancer, the constraint enforced in the last 10 years of the period curve prior to 1987 means that the small peak in incidence evident around 1984 in Figure 5.7 is not fully captured. This leads to the better projections for the spline restriction approach despite the shape projected by the spline drift approach actually being a better fit to the true shape.

Finally, Figure 5.9 contains the projections for Hodgkin's lymphoma for males. It is clear from the numbers on the y-axis, and the noisiness of the true number of cases that this is a rarer condition. Less common cancer sites provide even further issues when attempting to make projections as the chance of mis-fitting the observed data is increased. Even though the data are noisy, it is clear from Figure 5.9 that the downward trend in the number of cases between 1977 and 1987 is not continued beyond that point. This results in the spline restriction approach giving poor future projections of the total number of cases. However, it could be argued that

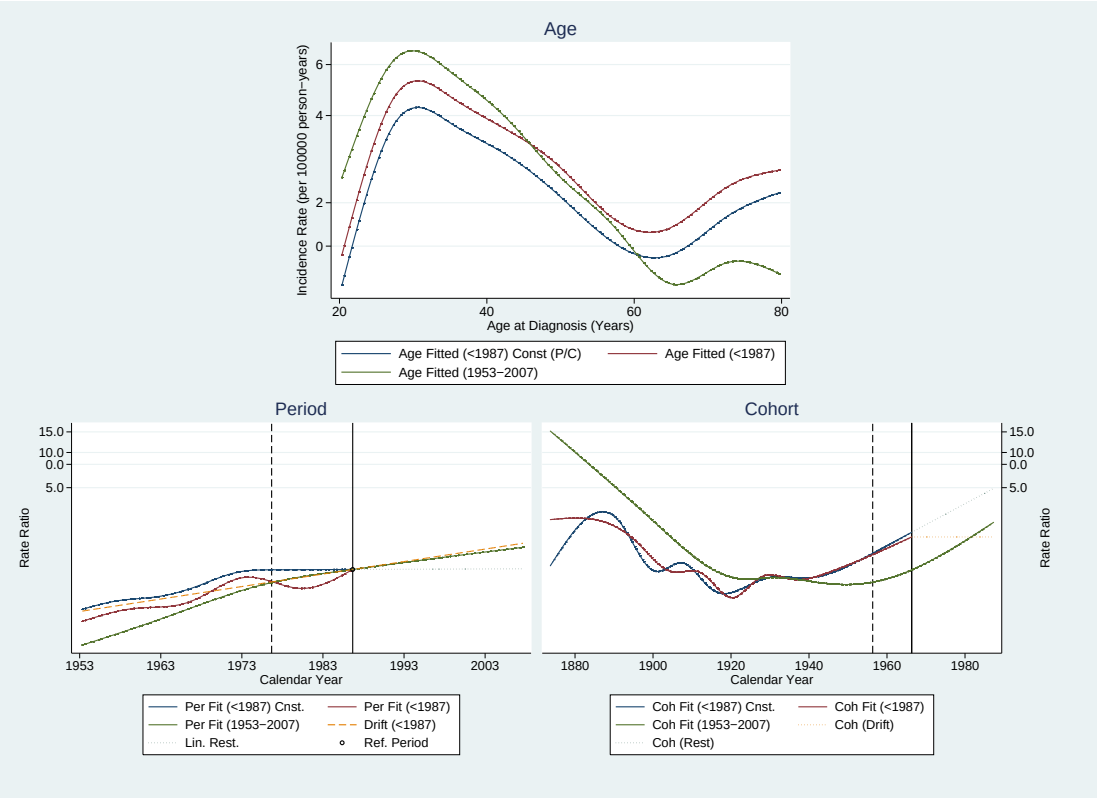

FIGURE 5.8. Age-period-cohort graph illustrating the two approaches. Testicular cancer for males.

the spline restriction approach appears to be making a more reasonable assumption based on the recently available data and that the full drift approach performs well by chance.

Figure 5.10 gives similar information to that expressed in Figure 5.9. This spline drift approach follows the full fitted trends for age, period and cohort better than the spline restriction approach for this example. Adding confidence intervals to this graph would highlight that there are fewer cases of Hodgkin's lymphoma than some of the other sites studied, as the confidence intervals are significantly wider for this site.

5.6. Discussion

A "case-by-case" approach to making projections for any given cancer site is more appropriate than applying a "one method fits all" approach. There can also be significantly different shapes for the incidence curves within a cancer site when splitting by gender. This is particularly the case for lung cancer in the examples shown. On this basis, the projections should

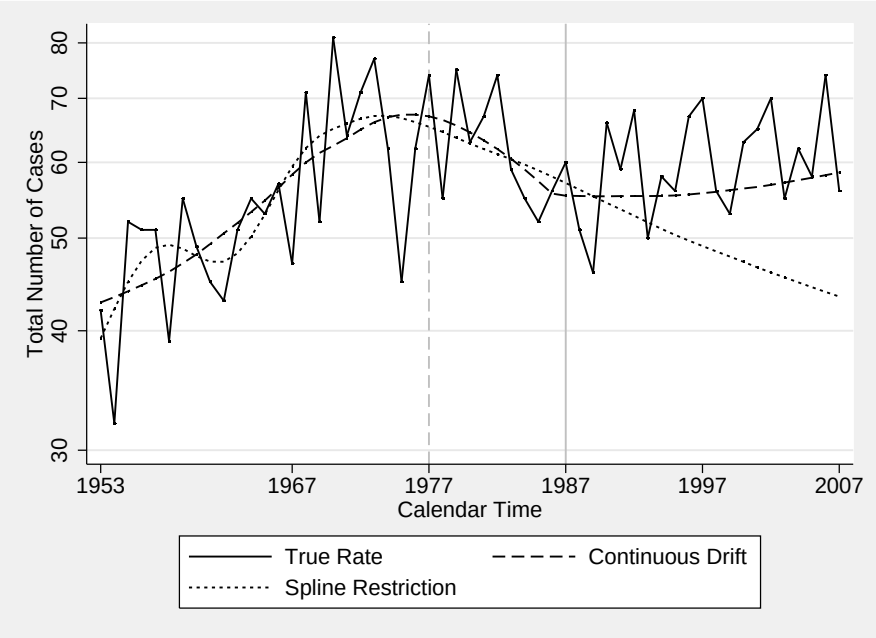

FIGURE 5.9. Projections from the two approaches from 1987. Hodgkin's lymphoma for males. A logarithmic link function was used for the projection models.

be made taking a considered approach and using external evidence where possible as a guide. Projections are often made for all cancer sites at the same time for nationwide evaluations of cancer trends. This usually leads to a single method of projection being applied universally. The results in this chapter highlight that this may not be the best course of action if a true estimate of future incidence is desired.

External information should be used as a guide to making projections and can also be used to decide whether or not the data available is appropriate to make projections. For example, if a screening program has recently been introduced this could have a large impact on the incidence rates. A second example is if constant disease reclassifications have been applied to the disease of interest, leading to a potentially confused shape to the incidence trends over time. For different countries, at different intervals in time, it is possible to select numerous examples of cancer sites where the above two statements will have an influence on the incidence data available. However, a careful, measured approach to projections using all of the information that is available, and appropriately stating the uncertainty in the estimates can still be of use to health planning authorities.

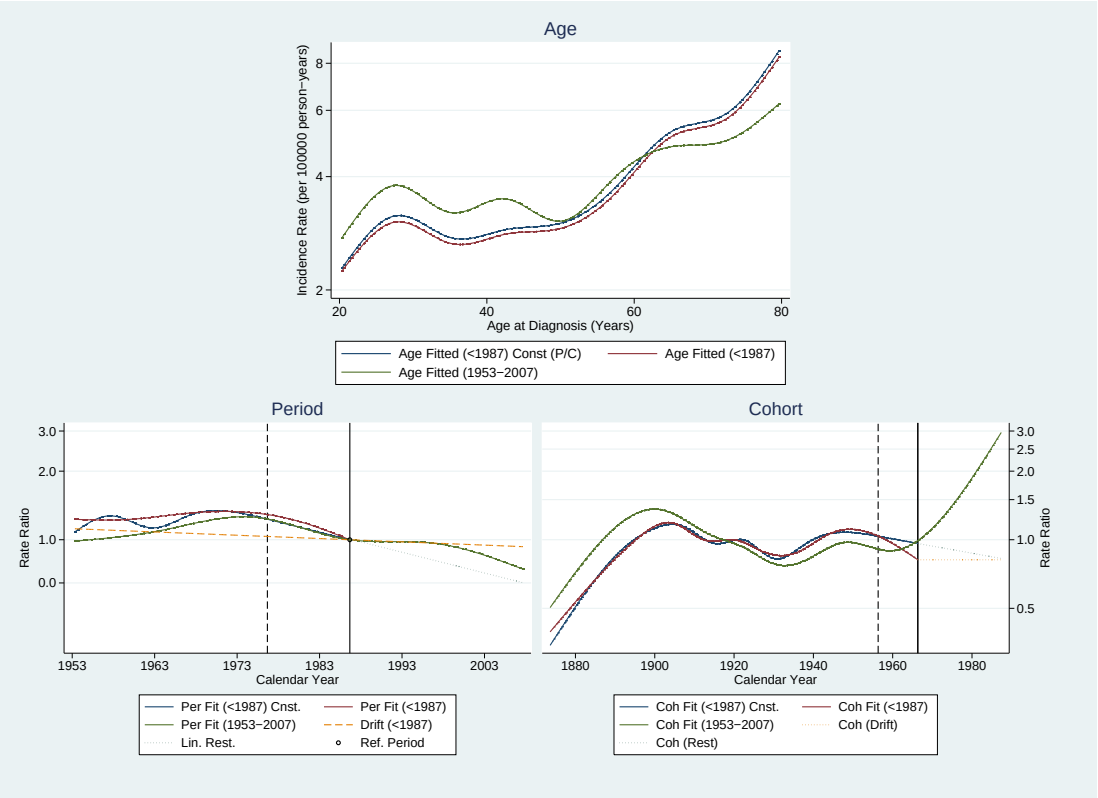

FIGURE 5.10. Age-period-cohort graph illustrating the two approaches. Hodgkin's lymphoma for males.

The plots showing the sensitivity to the boundary knot selection in the previous chapter can be of use for assessing the potential for the projection approaches to be uncertain, not just due to random variation but also due to the uncertainty involved in making a strong assumption as part of the projection. Further work is necessary in order to express the uncertainty clearly so that any projected estimates are understood to be a guide.

As a summary of the results and analyses from the previous two chapters, a list of guidelines for practice are given for incidence projections.

Recommendations for practice:

- Using a continuous representation of the data is more appropriate than using a factor model.
- Using recent data in order to make projections tends to perform better than using the long-term trend (drift).
- Looking at the observed data up until the point of projection can help to decide the most appropriate method to use. Plots of the age, period and cohort components as well as plots of the overall number of cases can both be of use.
- External evidence should be used as a guide for deciding the method of projection.
- In some extreme cases, it may be better not to make projections. For example, if a recent screening program has inflated the incidence figures that will be used for the projections.
- If projections are being made for a large number of sites there are bound to be sites where the projections given will be wrong.
- Cautious interpretation of any projected figures is always recommended.

CHAPTER 6

Survival Analysis

6.1. Chapter Outline

In this chapter, the concepts of survival analysis are introduced. Key recent advances in the analysis of survival data are explained so that they can be fully utilised in the following chapters, particularly methods for obtaining up-to-date and projected estimates of survival proportions.

6.2. Introduction

Survival analysis involves the analysis of data where interest lies in the time to a given event. In the context of cancer, the event of interest is often death or recurrence, and factors are investigated that either shorten or lengthen the time it takes for patients to experience that event. Cancer registries have been set up in most of the developed countries in the world. The cancer registries record data on individuals who are diagnosed with a case of cancer, and link the records to the death registries in order to ascertain the length of time each patient survives following their cancer diagnosis. The focus of this chapter will be on the analysis of registry data and will introduce the main tools of analysis used in this setting. One of the major approaches to analysing registry data is the use of methods to estimate relative survival. This concept will be introduced in this chapter before being further developed in the following chapter. A later chapter on survival analysis (Chapter 8) will concentrate in greater detail on the techniques necessary to obtain up-to-date, and potentially projected estimates of the proportion of people who survive following a cancer diagnosis.

Calculating the proportion of patients who are still alive at a given timepoint is vital for evaluating the proportion of people who currently are alive with a cancer diagnosis (the prevalence of the disease). The work given in the preceding chapters related to estimating the new cancer cases in a given period (the incidence). It is clear that both incidence and survival are needed to evaluate the current cancer burden.

6.3. Overall Survival

The simplest approach to analysing registry data is to look at the overall survival of the patients with a cancer diagnosis for any given cancer site. That is, where a death from any cause is treated as an event. The resulting estimate gives the proportion of patients that are still alive at a given timepoint after their diagnosis, with the point of diagnosis being used as the time origin (that is, t=0 at diagnosis). The calculation of this proportion is complicated by the fact that not all patients are followed up until their event time [Collett, 2003]. This is known as right censoring, because although the event time is unknown, we know that the patients' event time is to the right of the censoring time if considered on a timeline (that is, the event time will be after the point at which the patient had to be censored). Right censoring can occur for a number of reasons:

(i) Loss to follow-up - e.g. due to migration to another country meaning it is not possible to follow patients up until their event time.

(ii) Administrative censoring - this is more common for cancer registry data. At the time of analysis, there will be patients that are still alive having had a diagnosis of cancer. These patients are usually censored by choosing an appropriate cut-off point just prior to the analysis date.

Left censoring is another consideration, this is where the event of interest occurs before the time observed for a given patient. This is not a common occurrence when analysing cancer registry data as this information would be unlikely to be available in this setting. A final type of censored data is commonly referred to as interval censoring. This is when the event of interest is known to be after a certain date, and prior to a certain later date. That is, the event is known to fall within a certain interval. The fact that registry data is often recorded to the nearest month of diagnosis/death could be considered as a form of interval censoring. However, this information is not treated as interval censoring in the analysis of registry data; methods to deal with interval censoring are more common in other uses of survival analysis. An assumption that is made by the methods used to analyse registry data is that the right censoring is non-informative. That is, we assume that there is independence between the observation times and the censoring indicator (independent censoring). This means that it is assumed that there is no fundamental difference between the patients who are censored, and

those that are followed up completely. Consider two individuals with the same covariate pattern (same age, gender etc.) alive at time t with one having been censored before t and with one still under observation. Under independent censoring, both of these individuals should have equal chance of survival [Leung et al., 1997]. This assumption is usually considered to be valid for administrative censoring. However, there are arguments that censored observations due to loss to follow-up do not satisfy this independence assumption [Putter et al., 2007]. For example, it could be argued that those who emigrate are in a healthier state than those who do not; those who are less healthy may prefer to stay closer to their treatment centre. Fortunately, registry data from developed countries usually has a very low proportion of patients that are lost due to follow-up. This is particularly the case for the Finnish registry data used in the analyses conducted in this thesis due to the excellent facilities used for linking the cancer registry data and the death registry data.

6.3.1. Survival, hazard and probability density functions

The following definitions of the key quantities in survival analysis are given in [Collett, 2003]. The survival function, $S(t)$, is the probability that an individual will survive longer than time t. Let T be a continuous non-negative random variable denoting the time of occurrences for the event of interest. The survival function can then be expressed as:

$$S(t) = P(T \geqslant t). \tag{6.1}$$

The probability density function (pdf), denoted $f(t)$, is the probability of an event at time t and can be defined as:

$$f(t) = \lim_{\delta \to 0} \frac{P(t \leqslant T \leqslant t + \delta)}{\delta}. \tag{6.2}$$

The relationship between the survival function and the pdf is given as:

$$S(t) = \int_t^\infty f(u)\,\mathrm{d}u = 1 - \int_0^t f(u)\,\mathrm{d}u, \tag{6.3}$$

which implies (by differentiating both sides):

$$f(t) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -S'(t). \tag{6.4}$$

The hazard function, $h(t)$, is the instantaneous failure rate at time t. That is, the probability that the event will occur right at that instant given that the patient has survived until time-point t. This is defined as:

$$h(t) = \lim_{\delta \to 0} \frac{P(t \leq T \leq t + \delta | T \geq t)}{\delta}. \tag{6.5}$$

Under Bayes' theorem, these three functions satisfy the relation:

$$h(t) = \frac{f(t)}{S(t)}. \tag{6.6}$$

Using equations (6.4) and (6.6) implies:

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{\mathrm{d}}{\mathrm{d}t} \log(S(t)), \tag{6.7}$$

which implies (through exponentiation and integration) that;

$$S(t) = \exp\left(-\int_0^t h(u)\mathrm{d}u\right) = \exp\left(-H(t)\right), \tag{6.8}$$

where $H(t)$ is defined as the cumulative hazard function. The cumulative hazard function gives a measure of the sum of the exposure to the hazard up until time t. The (log) cumulative hazard scale is often used as the scale for modelling approaches, as is described in Section 6.4.3.

6.3.2. Non-parametric estimates

A simple non-parametric approach to estimating the overall survival function is the Kaplan-Meier estimate [Kaplan and Meier, 1958]. Let t_i denote the event times for a cohort of patients. The Kaplan-Meier estimator is then defined as:

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}, \tag{6.9}$$

where d_i define the number of events at time t_i, and n_i define the total number at risk at time t_i. This gives an estimate of the survival function using a step function which has a "step down" at each failure time. The Kaplan-Meier approach deals with censoring by allowing the denominator to be affected at the next failure time. The advantage of the Kaplan-Meier approach is that it does not assume a functional form the shape of the survival curve. However, when considering the survival of subgroups of patients, the functions can become unstable if there are only a small number of patients in any given subgroup.

The Nelson-Aalen estimator [Nelson, 1972; Aalen, 1978] is a non-parametric estimate of the cumulative hazard function, and is similar in principle to the Kaplan-Meier approach to estimating the survival function. The Nelson-Aalen estimator is given as:

$$\hat{H}(t) = \sum_{t_i<t} \frac{d_i}{n_i}, \tag{6.10}$$

where, as before, d_i define the number of events at time t_i, and n_i define the total number at risk at time t_i. This can be transformed to the survival function using the relation given in Equation (6.8). This will give a similar estimate to the Kaplan-Meier estimate for the survival function. The same arguments surrounding the interpretability of a step function apply to the Nelson-Aalen estimate. However, like the Kaplan-Meier estimate, a certain flexibility is allowed by not assuming a parametric form for the cumulative hazard function. Both of these approaches are defined heavily by the sample of patients that are under study, and are not conducive to making predictions for patients outside of the sample.

6.3.3. Survival or Mortality?

Both cancer survival and cancer mortality estimates can be provided from the population-based cancer registry data (provided information on the population size is also known). Cancer mortality is defined in a similar way to cancer incidence. Cancer mortality measures the number of new deaths due to cancer in a defined population within a specified time. Arguments have been made that because of biases in the collection of data for the time since diagnosis, mortality estimates should be used instead of survival figures for international comparisons [Beral and Peto, 2010]. However, cancer mortality estimates provide the cancer-specific mortality in the population within a specified time. Firstly, this measure assumes that the cause of death information is reliable. Also, the denominator for mortality rates is the entire population. Therefore, mortality rates are subject to both cancer incidence trends as well as trends in patient survival [Dickman and Adami, 2006]. In constrast, survival estimates are calculated within a cohort of patients that have had a diagnosis of cancer. The proportions calculated within this setting will not be influenced by trends in incidence in the same way.

6.3.4. Cox models: a semi-parametric approach

The Cox model [Cox, 1972] is a semi-parametric modelling approach to the analysis of survival

data. The Cox model allows the estimation of hazard ratios for particular covariates of interest and is used extensively to assess the impact that a given covariate has on the hazard function. The Cox model uses partial likelihood techniques to remove the need to give a parametric form to the baseline hazard. It is a standard assumption for a Cox model that the effect of a covariate is proportional over follow-up time. However, it is possible to relax this assumption.

The form of the Cox proportional hazards model can be expressed as:

$$h_i(t) = h_0(t) \exp(\boldsymbol{\beta}' x_i), \tag{6.11}$$

or on the log scale as:

$$\ln(h_i(t)) = \ln(h_0(t)) + \boldsymbol{\beta}' x_i. \tag{6.12}$$

The partial likelihood is used to estimate the parameters ($\boldsymbol{\beta}$) because the partial likelihood is independent of $h_0(t)$. This means that it is not necessary to estimate the baseline hazard whilst still being able to estimate the relative effect of each covariate. The ($\boldsymbol{\beta}$) parameters are often referred to as log hazard ratios. A hazard ratio can be interpreted in a similar way to the rate ratios calculated for the age-period-cohort models in Chapter 2. Both the Poisson model (for age-period-cohort models) and the Cox model are being used to model underlying rates. A hazard ratio expresses the effect of a given covariate on the hazard of an event. If the event of interest is considered to be death and the covariate is binary, then the hazard ratio gives the ratio of the mortality rates of one covariate level compared to the baseline covariate level. For a continuous covariate, the hazard ratio expresses the effect on the mortality rate for a unit increase in the covariate value.

6.3.4.1. *The Proportional Hazards Assumption*

The proportional hazards assumption is a strong assumption that is often made when applying the Cox model. Assuming proportional hazards means that only a single hazard ratio is used to summarise the entire time-scale and that over time, the effect of a covariate is assumed to be the same (on the hazard ratio scale - the underlying (baseline) hazard can still vary over time). This assumption is often invalid for population-based cancer studies.

It is possible to formally test the proportional hazards assumption after fitting a Cox model. There are also a number of graphical techniques for assessing whether or not the assumption is valid; the most common approach is to plot the Schoenfeld residuals against

follow-up time [Schoenfeld, 1982]. The Cox model has also been extended to allow for the case of non-proportional hazards by fitting a time-dependent effect for a given covariate. It is also possible to split the timescale into intervals in order to overcome the issue of having a time-dependent effect.

6.3.5. Reasons for Steering Away from the Cox Model

Although the Cox model is well established as a modelling framework for survival analysis, there are reasons for seeking an alternative modelling framework. Firstly, the Cox model is not ideally suited for dealing with the case of non-proportional hazards; that is, time-dependent effects. Time-dependent effects are very common in population-based cancer studies and a modelling framework that is suited to incorporating them easily would be preferable. Secondly, it would be desirable to consider a framework where other estimands used for the analysis of population-based data are easily available. For example, it is not easy to extend the Cox model to estimate relative survival. Relative survival is a key measure for population-based cancer studies and will be fully defined in Section 6.5. Finally, it is desirable to have a good estimate of the underlying (baseline) hazard in order to give estimates of both relative and absolute measures of risk. Standard alternatives to the Cox model use a parametric form for the baseline hazard; such as the exponential, Weibull or gamma distribution. A criticism of these parametric models is that they are not flexible enough to capture the underlying shape of the hazard in the majority of scenarios.

In the next section, a modelling framework is introduced that is superior to the Cox model in respect to the above criticisms whilst also having the flexibility to capture the underlying shape of the hazard parametrically.

6.4. Flexible Parametric Models

6.4.1. Introduction

There are a number of long-established methods for modelling cancer patient survival from population-based cancer registry data. To find a model that has a similar framework to the age-period-cohort model would be of great benefit when trying to combine the models to assess the cancer burden (see Chapter 9). The methods that have been suggested for obtaining prevalence estimates vary in their use of survival estimates. Some of the techniques use estimates of overall

survival [Heinävaara and Hakulinen, 2006] whereas others use estimates of relative survival [Capocaccia and De Angelis, 1997]. It would, therefore, be useful to have a technique whereby these two estimates can be obtained from the same modelling framework. A method of analysis that satisfies both of the above criteria is the use of flexible parametric modelling for the analysis of survival data. Flexible parametric models use restricted cubic splines to model the shape of the underlying baseline hazard and allow a parametric estimation of the model parameters. Flexible parametric models are capable of estimating both overall, and relative, survival in the same framework providing that the appropriate data is available for the expected mortality rates in the reference population. It is also possible to estimate cause-specific survival in the flexible parametric modelling framework provided that information on cause of death is available.

6.4.2. Background

Flexible parametric models were first introduced by Royston and Parmer [Royston and Parmar, 2002]. The software to carry out these types of models has recently been improved and updated [Lambert and Royston, 2009] with better capability of modelling time-dependent effects. The flexible parametric model offers a parametric alternative to the Cox proportional hazards model for survival data. The flexibility of the models is introduced through using splines to capture the shape of the baseline hazard.

The flexible parametric models provide a framework in which the modelling of a simple proportional hazards model, and a much more complex time-dependent effects model can be easily fitted. The models also allow both overall and relative survival [Nelson et al., 2007] to be fitted in a very similar way. These models can also be used to carry out a period analysis (a method for obtaining up-to-date estimates of survival; see Section 6.6) provided that the data is set up appropriately because delayed entry can be incorporated. Each of these methods can equally be carried out using Cox regression. However, the methodology and practice can be much more complex when using Cox regression.

The fact that the flexible models described [Royston and Parmar, 2002] are fitted parametrically means that the model parameters can be transformed to express the differences between the groups in numerous ways [Lambert et al., 2011]. For example, it is possible to quantify the difference in the survival proportion between groups or estimate the difference in mortality rate between two groups. These differences are absolute measures, which contrasts to the usual

relative measures (hazard ratios) commonly associated with survival analysis. The absolute differences are achievable due to the full modelling of the baseline hazard function. These measures lead to a clearer understanding of the underlying risk and help to quantify what the differences actually mean to patients [Lambert and Royston, 2009].

The flexible parametric models are performed by modelling on the log cumulative hazards scale rather than, the more standard, log hazard scale. Under proportional hazards, this does not affect the estimation of the hazard ratios as they are equivalent on both scales. The flexible parametric models are simply an extension to the simple Weibull model that would be fitted on the log cumulative hazard scale. The Weibull model is well-known to be inflexible, and relatively poor at capturing the shape of the hazard function due to the monotonic shape that is imposed. It is possible to relax the linearity that the Weibull model introduces on this scale by replacing the linear term with a collection of restricted cubic spline terms. The basic mathematical details are given in the following section; further details can be found in the cited literature [Royston and Parmar, 2002; Lambert and Royston, 2009; Lambert et al., 2010].

6.4.3. Modelling on the log cumulative hazard scale

The flexible parametric model formulation is an extension of a more simple Weibull model for the survival curve. The Weibull model for $S(t)$ can be expressed as:

$$S(t) = \exp\left(-\lambda t^{\gamma}\right), \tag{6.13}$$

where λ and γ are real, positive values to define the shape and scale of the curve.

By taking the natural logarithm of equation (6.8) we have;

$$H(t) = \left(-\ln(S(t))\right), \tag{6.14}$$

If we introduce the Weibull survival curve from equation (6.13) and again take the logarithm to convert to the log-cumalative hazards scale we have:

$$\ln\left\{H(t)\right\} = \ln\left\{\left(-\ln\left(\exp\left(-\lambda t^{\gamma}\right)\right)\right)\right\} = \ln\left\{\lambda t^{\gamma}\right\} = \ln(\lambda) + \gamma \ln(t). \tag{6.15}$$

It is clear that equation (6.15) shows that on the log cumulative hazard scale, the Weibull model is a linear function for $\ln(t)$, with intercept $\ln(\lambda)$ and gradient γ. It is possible to introduce

covariates into this model, with $\mathbf{x}$ being a matrix for the covariate values for each patient, x_{ij}, and $\boldsymbol{\beta}$ being the vector of the j coefficients for the model:

$$\ln\{H(t|\mathbf{x})\} = \ln(\lambda) + \gamma\ln(t) + \mathbf{x}\boldsymbol{\beta}. \tag{6.16}$$

On this scale, it is easy to envisage making this model more flexible by using splines to model the $\ln(t)$ component rather than using a linear function of $\ln(t)$. The flexible parametric approach [Royston and Parmar, 2002] uses restricted cubic splines to introduce this flexibility. Restricted cubic splines were introduced and described in detail in Section 2.8.1. The same formulation is used here, and the spline function will be denoted by $s\left(\ln(t)|\gamma, \mathbf{k_0}\right)$, where $\mathbf{k_0}$ are the selected knots for the baseline spline function. The subscript of 0 is used to distinguish between the knot selection for the baseline spline function, and any knot selection that may be used for the effect of continuous covariates or for the introduction of time-dependent effects.

In the case of the flexible parametric proportional hazards model we therefore have:

$$\ln\{H(t|\mathbf{x})\} = s\left(\ln(t)|\gamma, \mathbf{k_0}\right) + \mathbf{x}\boldsymbol{\beta} = \eta. \tag{6.17}$$

The linear predictor, η, gives the log-cumulative hazard function. Therefore, using the relation in Equation (6.8), the survival function is given by:

$$S(t) = \exp\left(-\exp(\eta)\right). \tag{6.18}$$

Also, using Equation (6.7), the hazard function is obtained by:

$$h(t) = \left[\frac{\mathrm{d}}{\mathrm{d}t}s\left(\ln(t)|\gamma, \mathbf{k_0}\right)\right]\exp(\eta). \tag{6.19}$$

The advantage of modelling on the log cumulative hazard scale is that these quantities ($S(t)$ and $h(t)$) are directly estimable without the need for numerical integration techniques. Numerical integration would be required if the model was applied on the log hazard scale in order to estimate the cumulative hazard function. The spline function is a collection of cubic polynomial terms. Therefore, the derivative of the spline function required in Equation (6.19) can be calcuated directly and easily.

6.4.4. Simple extension to non-proportional hazards

A nice feature of the flexible parametric approach is the simplicity in which it can be extended to the case where the effect of a covariate is dependent on the point in the timescale; this is often referred to as a time-dependent effect or non-proportional hazards.

Equation (6.17) can be modified in the case of time-dependent covariates. If we have D time-dependent effects, where D is less than or equal to the total number of covariates, then we have:

$$\ln\{H(t|\mathbf{x})\} = s\left(\ln(t)|\gamma, \mathbf{k_0}\right) + \sum_{j=1}^{D} s\left(\ln(t)|\delta_j, \mathbf{k_j}\right) x_{ij} + \mathbf{x}\boldsymbol{\beta}. \tag{6.20}$$

The terms with time-dependent effects now have a defined interaction with follow-up time. This allows the effect of the covariate to vary over the time period. The flexibility of the above approach also means that each time-dependent effect can in theory have a different number of knots for the spline terms. In the original literature for flexible parametric models [Royston and Parmar, 2002; Nelson et al., 2007] the same knot placements were used for the baseline hazard and the time-dependent effects, which led to cases of overfitting. The formula above allows different knot placements for the time-dependent effects [Lambert et al., 2010], which can allow a more parsimonious time-dependent model to be fitted. It is often the case that the complexity required for the underlying hazard is greater than that needed for the departures from proportionality required for the time-dependent effects. This is similar in principle to the "reduced set of splines" used in Chapter 2 for the interaction between gender and age in the APC model setting.

6.4.5. Example

The data used in this example relate to colon cancer diagnoses in Finland from the early 1950s through to the end of 1990s, with follow-up until 2008. This example is designed to give a simple indication of the use of flexible parametric models for estimating overall survival. Further details of the scope and flexibility of the method are given in the associated literature [Royston and Parmar, 2002; Lambert and Royston, 2009] and in the later chapters (Chapters 7, 8 and 9). The periods of diagnoses were grouped into decades of diagnosis, and the age at diagnosis was categorised into 3 age groups (< 45, $45-60$, $60+$). Five degrees of freedom were

used for the underlying baseline hazard and a flexible parametric model was fitted to the colon data with the categorised age and period terms included in the model.

Table 6.1 gives the parameter estimates from the flexible parametric model assuming proportional hazards. The hazard ratio for the youngest age group (< 45) is 0.5710. This means that, compared to the middle age-group ($45 - 60$), the mortality (hazard) in the youngest age-group is around 43% lower. Using the same dataset, it is possible to compare the results obtained from fitting a Cox model to those obtained from the flexible parametric approach. The estimated hazard ratios for the equivalent Cox model are contained in the final column of Table 6.1. It is clear that the estimates are broadly similar from both methods. The extra significant figures given here are purely for comparison and it is unlikely that reported figures would be given to such accuracy. The standard errors for the hazard ratios are also reported in brackets for both of the measures. Again, the estimates obtained from both approaches of very similar.

Covariate Level	**Flexible Parametric HR (SE)**	**Cox HR (SE)**
Age < 45	0.5710 (0.0194)	0.5700 (0.0194)
Age 45-60	1.0000 (-)	1.0000 (-)
Age 60+	2.0338 (0.0361)	2.0361 (0.0362)
Period 1950s	1.8143 (0.0483)	1.8072 (0.0481)
Period 1960s	1.3423 (0.0296)	1.3408 (0.0295)
Period 1970s	1.0000 (-)	1.0000 (-)
Period 1980s	0.8148 (0.0147)	0.8151 (0.0147)
Period 1990s	0.6888 (0.0121)	0.6924 (0.0121)

TABLE 6.1. Comparison of HRs from Cox model and a flexible parametric model with 5 degrees of freedom for the baseline.

Figure 6.1 shows the resulting survival estimates from the flexible parametric model for the middle age group; those that were diagnosed between the ages of 45 and 60. The plotted predictions are under the assumption of proportional hazards for the effect of decade of diagnosis. Figure 6.1 shows that the overall survival for patients improves in each decade from the 1950s up until the 1990s. The increases in the proportion of patients that are still alive after 15 years are likely to be due to improvements in treatment. However, it could also be the case that patients are being diagnosed at an earlier point in their development of cancer during the later decades, which can lead to an artificial increase of a patient's survival time. This is commonly

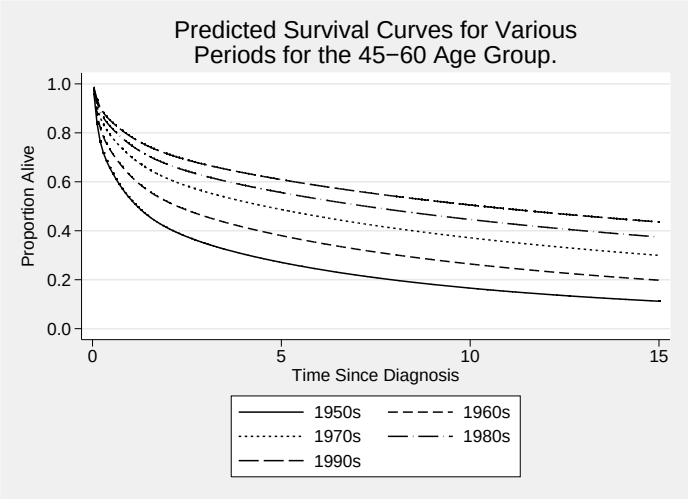

FIGURE 6.1. Overall survival from colon cancer using the flexible parametric models. The lines indicate different decades of diagnosis.

referred to as lead time bias [Hutchison and Shapiro, 1968; Morrison, 1982] (see Figure 6.5 and the corresponding description).

These results are based on broad categorisations for age, and period of diagnosis. These models can be fitted using splines to give a continuous representation for covariates of interest. It is these types of models that will be used in the estimation of prevalence in Chapter 9. It is fairly unrealistic to assume that a patient's survival makes a discrete jump depending on the decade of diagnosis, and that a more realistic scenario is that there is a more continuous development of improved survival over calendar time. The flexibility of these models means that spline terms can be fitted for each covariate whilst also using splines to capture the underlying baseline hazard.

6.4.6. Wider use of flexible parametric models

The flexibility of the flexible parametric models goes beyond the "flexible" capturing of the baseline hazard. These models can be used in a wide-range of analyses that are carried out on time-to-event data for population-based studies. They also have applications for the analysis of randomised controlled trials data [Albain et al., 2010; Gore et al., 2010], and in other more diverse settings [Johansson et al., 2011].

6.5. Relative Survival

6.5.1. Introduction

Relative survival has become an increasingly common measure in population-based data analysis. The reason for this is because relative survival is useful for comparisons between groups in that it is adjusted for the fact that different populations may have different levels of background risk of death. Relative survival methods are used to try to obtain an estimate of net survival; that is, the probability of surviving the disease of interest in the absence of death from other causes. This is a hypothetical measure in that we are trying to estimate the proportion of deaths due to cancer in a "world where you cannot die of other causes". The difference between relative and overall survival is the need to calculate the background mortality in an appropriate reference population, which, for most cancer sites, is assumed to be the general population of the country or region. Therefore, the expected (or background) mortality is normally obtained from nationwide or regional population mortality figures, which are obtained in yearly intervals for age and calendar time, whilst also being estimated separately for each sex. Relative survival as a function of time ($R(t)$) is defined as:

$$R(t) = \frac{S(t)}{S^*(t)}, \tag{6.21}$$

where $S^*(t)$ is the background survival in the population, and $S(t)$ is the observed survival for the cancer patients. This can also be written on the hazard scale, with the excess hazard (mortality), $\lambda(t)$, defined by:

$$\lambda(t) = h(t) - h^*(t), \tag{6.22}$$

where $h(t)$ is the observed hazard (mortality) amongst the cancer patients, and $h^*(t)$ is the background hazard (mortality) for a relevant comparative population. This shows how the total mortality, $h(t)$, is split into two components; the background mortality due to other causes, $h^*(t)$, and the excess mortality associated with the disease of interest, $\lambda(t)$.

6.5.2. Life-table Approaches

There are three life-table approaches to estimating relative survival that are well established and commonly used. The three life-table approaches (Ederer I [Ederer et al., 1961], Ederer II

[Ederer and Heise, 1959], and the Hakulinen method [Hakulinen, 1982]) differ purely in their estimation of $S^*(t)$ when calculating relative survival using equation (6.21). The time-scale is usually split into yearly intervals, with separate estimates made for each time interval. These methods have been commonly used for the estimation of relative survival for decades, and are generally applied to coarsely grouped data (often yearly intervals of follow-up time) and were designed with ease of computation in mind. A detailed description of these approaches will be given in the following chapter.

6.5.3. Pohar-Perme et al. approach.

A recent publication [Perme et al., 2011] has shown a new estimator of relative survival that claims to estimate the true net survival provided an appropriate life-table is used. The approach uses a weighting approach that inflates the group at risk in order to account for those that have been lost due to deaths from other causes to obtain an unbiased estimator of the net survival. There is detailed discussion in the paper about the alternative life-table approaches and the fact that the Ederer II approach estimates the "observable net survival", rather than the net survival itself as it does not appropriately deal with the fact that the censoring of patients who die from other causes is informative [Perme et al., 2011].

In the paper, the definition of the overall average net survival for the cohort of patients, $S_N(t)$ as a whole is given as:

$$S_N(t) = \frac{1}{n}\sum_{i=1}^{n} R_i(t). \tag{6.23}$$

That is, the estimated net survival is the average of the individual-level relative survival estimates. Appropriately estimating $R_i(t)$ requires consideration to be given to the key covariates that influence relative survival. This is considered further in Chapter 7.

6.5.4. Flexible Parametric Model for Relative Survival

There are also a number of modelling approaches to estimating relative survival [Estève et al., 1990; Dickman et al., 2004; Nelson et al., 2007]. A detailed description and comparison of the available estimates are available in the next chapter, Chapter 7. However, it should be noted at this point that the flexible parametric approach can be extended to the estimation of relative survival [Nelson et al., 2007]. This extension involves incorporating the background rate of death in the general population so that the models apply instead to the cumulative

excess hazard scale. The background hazard only needs to be estimated at the event times, so the method still does not require splitting the timescale.

Integrating Equation (6.22):

$$H(t) = H^*(t) + \Lambda(t), \tag{6.24}$$

where $H(t)$ is then the overall cumulative hazard, $H^*(t)$ is the cumulative expected hazard, and $\Lambda(t)$ is the cumulative excess hazard. The extension to relative survival for the flexible parametric approach is then to simply model on the log cumulative excess hazard scale (Similarly to Equation (6.17)):

$$\ln\{\Lambda(t|\mathbf{x})\} = s\left(\ln(t)|\gamma, \mathbf{k_0}\right) + \mathbf{x}\boldsymbol{\beta} = \eta. \tag{6.25}$$

In order to make this extension the information from life-tables is used to estimate the cumulative expected hazard.

6.5.5. Example

Table 6.2 compares the results obtained in Section 6.4.5 to those that are obtained from a flexible parametric relative survival model. Exactly the same covariates are included in the relative survival model. However, the relative survival model takes into account the background population mortality information for each patient. Therefore, the estimates obtained from the flexible parametic relative survival model are excess hazard ratios; they relate to the excess mortality associated with a diagnosis of cancer.

Covariate Level	**Observed Survival HR (SE)**	**Relative Survival Excess HR (SE)**
Age < 45	0.57 (0.0194)	0.65 (0.0241)
Age 45-60	1.00 (-)	1.00 (-)
Age 60+	2.03 (0.0361)	1.52 (0.0313)
Period 1950s	1.81 (0.0483)	1.94 (0.0585)
Period 1960s	1.34 (0.0296)	1.42 (0.0369)
Period 1970s	1.00 (-)	1.00 (-)
Period 1980s	0.81 (0.0147)	0.74 (0.0172)
Period 1990s	0.69 (0.0121)	0.60 (0.0137)

TABLE 6.2. Observed vs Relative survival using a flexible parametric model with 5 degrees of freedom for the baseline.

Looking at Table 6.2, the difference between the two columns are that one contains estimates of hazard ratios in the observed survival setting, whereas the other contains estimates of excess hazard ratios in the relative survival setting. The excess hazard ratios for the effect of age are less extreme than the hazard ratios in the observed survival setting. This is due to the fact that the background mortality is strongly affected by age and this effect is included in the overall survival analysis. The excess hazard ratio indicates that cancer-specific mortality is also associated with age, but less strongly. The oldest age-group (60+) have an excess mortality rate that is 1.52 times higher than that of the middle age-group (45-60). This is a proportional excess hazards model, so this ratio is assumed to be the same at each point across follow-up time. Table 6.2 also indicates that the effect of decade of diagnosis is a stronger effect for the relative survival model, compared to the overall survival model. This highlights that decade of diagnosis is more strongly associated with the excess mortality due to cancer than with the overall mortality rate. It should be noted that all of the estimates contained in Table 6.2 are relative measures of risk.

6.5.6. Cause-Specific vs Relative Survival

Another method for estimating net survival would be to use cause of death information to categorise each death. The deaths need to be categorised into two seemingly simple categories; "death caused by the cancer of interest" and "death not caused by the cancer of interest". However, this is not a simple process from the information recorded on a death certificate. Imagine a scenario whereby a patient diagnosed with cancer has committed suicide. Can it be certain that this is independent of the cancer diagnosis that this patient has received? And, contrary to that, can it be certain that this patient committed suicide due to their diagnosis? Although an extreme example, there are other more orthodox examples, such as death due to cardiovascular disease that may well have been caused by the treatment that the patient has been prescribed. It is for this reason that it is difficult to fully classify whether or not a death is, or is not due, to the cancer of interest [Begg and Schrag, 2002].

A secondary issue with the cause-specific analysis is that the cause of death information recorded on the death certificate is often inaccurate or poorly recorded [Flanders, 1992; Satariano et al., 1998]. The relative survival approaches described in the previous section circumvent the need for death certificate information by comparing the survival experience in the cohort

of patients with cancer to the general population. In this way, we get the estimate of excess deaths associated with the cancer of interest, which is the quantity that we wish to estimate.

Therefore, in the analyses carried out in this thesis, attention will be given to methods using relative, rather than cause-specific, survival. Further details on the comparison of the two approaches have recently been given [Sarfati et al., 2010].

6.6. Period Analysis

One method that has become increasingly popular for obtaining up-to-date estimates of survival proportions is period analysis [Brenner and Gefeller, 1996; Brenner et al., 2004b]. Period analysis uses the survival experience of recently diagnosed patients to give a more accurate estimate of the current survival proportion for patients in the period of interest. This is performed by defining a window, or period, of interest and then using delayed entry techniques by left truncating the survival times for patients that were diagnosed prior to the start of the window. The benefit of this method is that short-term survival estimates are only dictated by those patients diagnosed more recently. Figure 6.2 gives an illustration of how the defined window affects the cohort of patients, and the survival times that are used in the period analysis. The period window is indicated by the dashed vertical lines. The dotted horizontal lines portray the survival experience for the patients prior to the start of the window. The solid lines are the survival experience used in the period analysis; in a traditional cohort approach the entire survival experience would be used for each patient. Those surviving for a short time that fall outside of the window are not included in the analysis (See Patient 1 in Figure 6.2). This means that more recent data is used to define the proportion surviving for the early years of follow-up and less relevant (older) data about the short-term survival is excluded. If short-term survival improves over time, the period estimate will lead to a higher estimate of short-term survival (and consequently long-term survival, as it is a cumulative measure). Period analysis has been shown to give better estimates of survival empirically in a number of settings [Brenner and Hakulinen, 2002; Brenner, 2003].

Period analysis can be performed in any technique of estimating survival provided that the technique has the capability of incorporating delayed-entry. This includes the flexible parametric modelling framework introduced in Section 6.4.

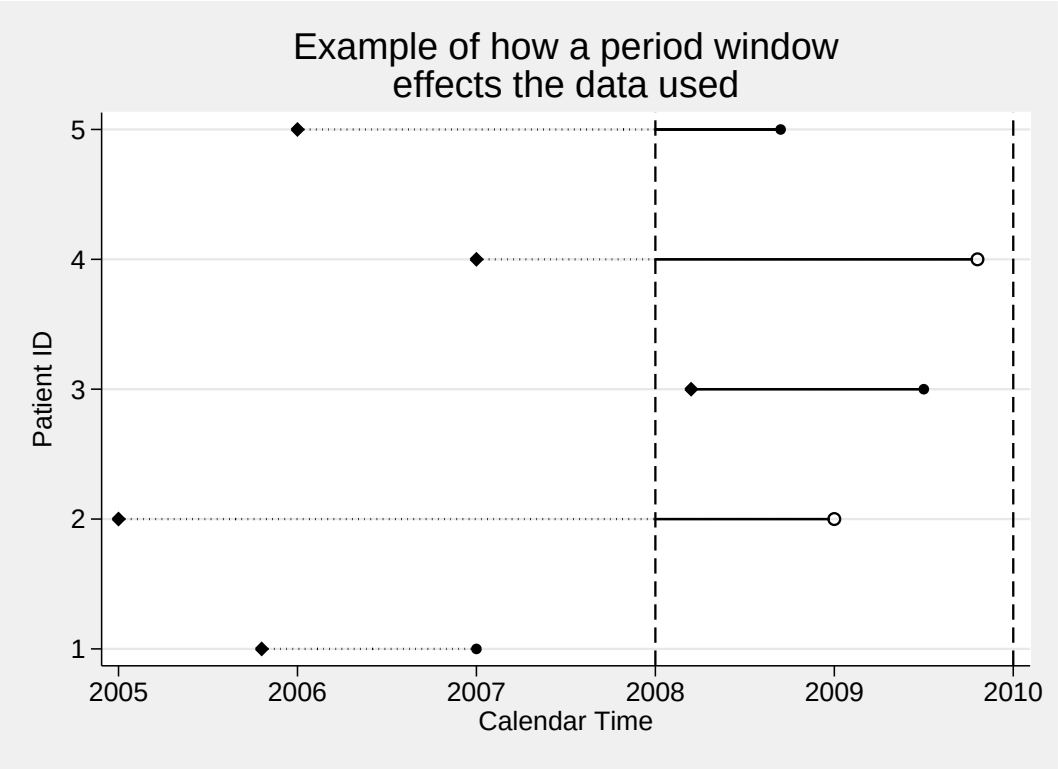

FIGURE 6.2. Example of a period window from 2008 to 2010.
The period window is indicated by the dashed vertical lines. The dotted horizontal lines show the survival experience for the patients before the start of the window. The solid lines are the survival experience used in the period analysis (with appropriate delayed entry). A filled in circle indicates a death, whereas a censored variable is indicated by a hollow circle. The diamonds indicate the date at diagnosis.

6.7. Modelled Period Analysis

Modelled period analysis is becoming a more widely used method for obtaining up-to-date, and potentially projected, estimates of relative survival [Brenner et al., 2009a,b; Pulte et al., 2010]. The basic priciple behind the approach involves fitting a linear trend for attained calendar year within a GLM in order to attempt to capture the improvements over calendar time [Brenner and Hakulinen, 2006a, 2008, 2009]. It is thought that this will give appropriate up-to-date estimates of relative survival. It is a simple extension to period analysis which was introduced in the previous section.

It is felt that flexible parametric models can be used in order to capture and project the trend in a similar way to the method proposed by Brenner *et al.* In Chapter 8, modelled period

analysis using the flexible parametric framework will be explained and used in a retrospective analysis of Finnish cancer registry data.

6.8. Models for Cure

6.8.1. Introduction

A number of modelling techniques have been proposed to estimate the level of statistical cure that a population of cancer patients experience. Statistical cure refers to the point at which the patients are at no excess risk of death from their cancer diagnosis compared to an appropriate reference population; often the mortality rates for the general population are used [Lambert et al., 2007]. The modelling techniques that are used define the time of cure at infinity and allow an estimate of the proportion of the cancer patients that are cured to be obtained; often referred to as the cure fraction. Further to this, a survival distribution for the "uncured" group (or those that are "bound to die") can be retrieved from the models. This survival distribution allows estimates of the time it takes for any given proportion of the "uncured" group to die. This, along with the estimate of the cure fraction, gives a useful statistic for comparing across time periods to assess improvements in cancer patient treatment and care.

6.8.2. Mixture Models for Cure

One of the modelling techniques that has been applied to cure models are known as mixture cure models. As mentioned, these models make the assumption that the time to statistical cure is defined at infinity. The mixture cure models also make the assumption that the patients can be split into two groups; those that are cured and those that are "bound to die", at time zero.

The survival function for the mixture model takes the form:

$$S(t) = S^*(t)(\pi + (1 - \pi)S_u(t)), \tag{6.26}$$

where $S^*(t)$ is the expected survival in the reference population, π is the proportion cured (cure fraction), $S_u(t)$ is the survival distribution for the "uncured" group and $S(t)$ is the overall survival rate. For more details of the model specifications the reader is referred to the cited literature [De Angelis et al., 1999; Lambert, 2007; Lambert et al., 2007]. Fitting this model to the survival data for a group of patients will yield a predicted value for the proportion of the patients that experience statistical cure, π. A parametric form is often specified for the

survival of the uncured, such as the exponential, Weibull, log-normal or gamma distribution. In practice, the Weibull distribution is usually used [Verdecchia et al., 1998; Shah et al., 2008; Eloranta et al., 2010].

6.8.3. Non-Mixture Models for Cure

A second modelling techniques that has been applied to cure models are known as non-mixture cure models [Sposto, 2002; Lambert et al., 2007]. These models also apply a function that has an asymptote at the cure proportion, π. The survival function for the non-mixture model takes the form:

$$S(t) = S^*(t)\pi^{F_z(t)}, \tag{6.27}$$

where $F_z(t)$ is a distribution function. As with the mixture cure model a parametric form is often selected for $F_z(t)$ and it is common to use a Weibull distribution. The non-mixture model can be re-expressed as a mixture cure model:

$$S(t) = S^*(t)\left(\pi + (1-\pi)\left(\frac{\pi^{F_z(t)} - \pi}{1-\pi}\right)\right), \tag{6.28}$$

which allows the estimation of the survival experience of those "bound-to-die", as well as an estimate of the cure proportion.

6.8.4. Example

Using the Finnish Cancer Registry data [Finnish Cancer Registry] for colon cancer from 1953 until the end of 2008, it is possible to give a brief example as to how the mixture cure models perform and the types of results that can be obtained from these models. Cure models can be applied using a user-written command [Lambert et al., 2007] in Stata [StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP]. To simplify the resulting graphs, the variables relating to the age of the patients and the periods for the times of diagnosis are categorised (as in section 6.4.5). In this example, the Finnish colon data can be used to compare the values obtained for the cure models across different decades to try to quantify improvements in both treatment and care. A mixture model using a Weibull distribution for the survival times has been applied to the data.

Figure 6.3 shows the results of the mixture cure model for a specific age-group and period. The graph relates to patients aged between 45 and 60, and for those that were diagnosed

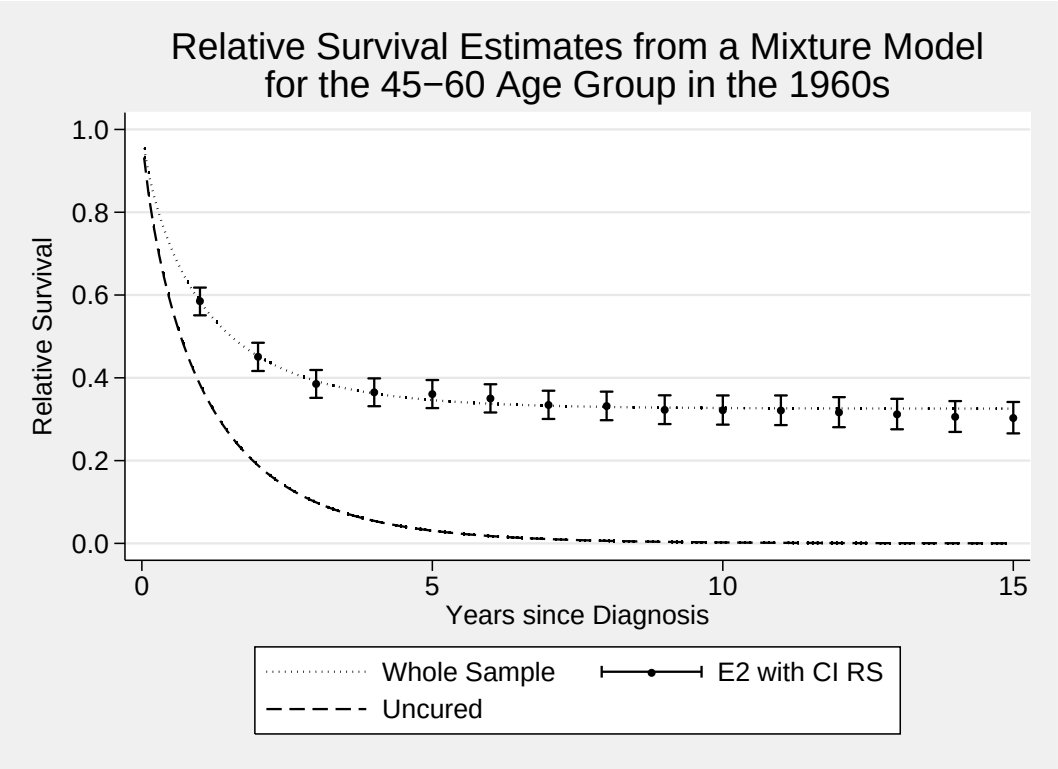

FIGURE 6.3. Results of the mixture cure model for the 1960s.

during the 1960s. The plot gives the relative survival against the survival time, in years. The dashed line relates to the whole sample of patients and shows the relative survival reaching a plateau beyond a certain point in survival time. This plateau is forced by the specification of the mixture cure model, so to check that this relates to the observed data, the Ederer II estimates introduced in Section 6.5.2 are overlaid on the plot. It is clear that there is fairly good agreement between the cure model and the estimates. The plateau indicates that the excess mortality is approaching zero and that the survival in the cancer patients has reached the same level as would be expected in the reference group; which in this case is the general population of Finland. The cure model estimates the value of relative survival for the entire population for the time$= \infty$. This gives the estimated proportion of the population that are cured, which is also known as the cure fraction. In this case we can see that the value is 0.33; that is, around 33% of the cancer patient population can be assumed to be statistically cured. The dashed line on the curve gives the distribution of the survival times for those that are "uncured" or "bound to die". It is clear that the majority of those patients who die from their colon cancer in this

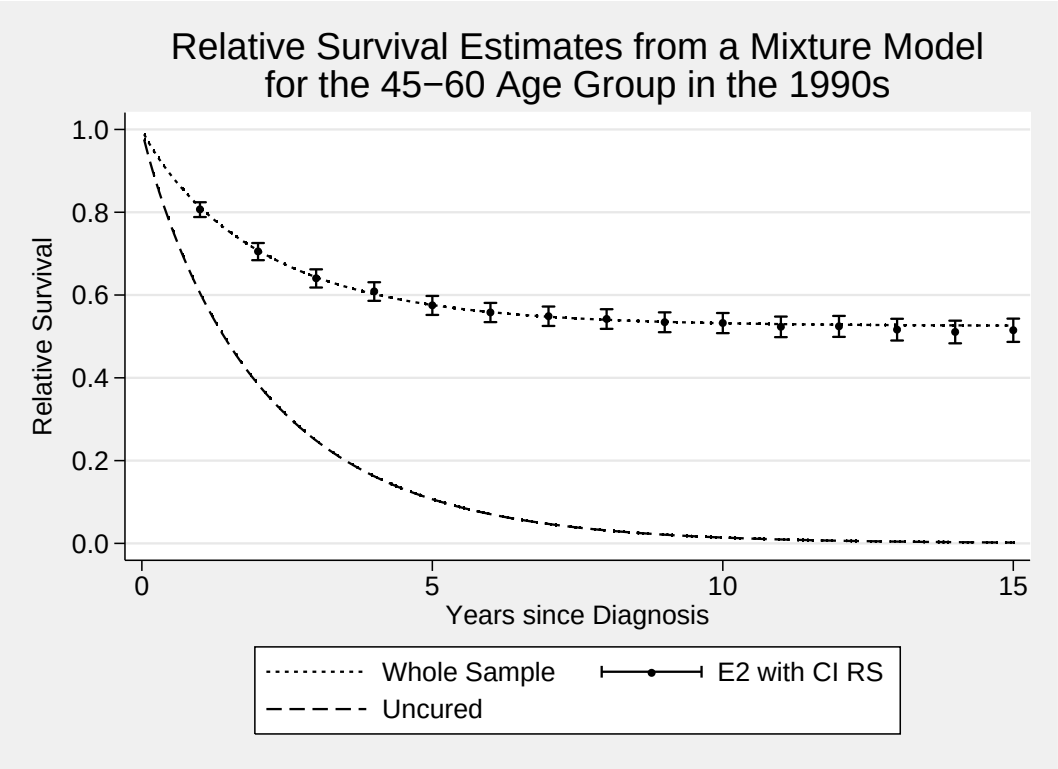

FIGURE 6.4. Results of the mixture cure model for the 1990s.

age group in the 1960s have died by around 5 years after their diagnosis. However, it is worth noting that because a relative survival model has been fitted, this definition is true only in the hypothetical world where cancer is the only potential cause of death.

Figure 6.4 gives the results of the mixture model for the 1990s decade for the same age group (45-60 years old). The interest lies in comparing the two curves from each of the graphs. It can be seen that the dotted line relating to the entire sample of the patients plateaus at a much larger value when comparing Figure 6.4 to Figure 6.3. The proportion cured in the 1990s is 0.53. This indicates that there has been a large increase in the proportion of the patients that are cured when comparing the patients diagnosed in the 1960s to those that are diagnosed in the 1990s. This increase is indicative of improvements in the treatment of colon cancer in Finland. The other consideration to be made when comparing the two graphs is the different shapes that are evident for the survival distributions of those that are "bound to die". The difference in the shape of the dashed lines indicates that those patients that are "bound to die" diagnosed in the 1990s tend to live longer than those that are "bound to die" but diagnosed in the 1960s.

This could be due to improvements in the care and treatment of those patients that are "bound to die" from their cancer. However, these values can be susceptible to bias; particularly the issue of lead time bias [Hutchison and Shapiro, 1968; Morrison, 1982]. Lead time bias relates to the lag that occurs from the onset of cancer to the actual diagnosis. The issue is illustrated in Figure 6.5. A diagnosis that happens more quickly after the actual onset of cancer could lead to an artificially inflated survival time compared to a patient diagnosed later but with the same true survival time (that is, the time from onset until death). For example, this lag could be reduced in the case of a newly-introduced screening program, which would then have the effect of artificially increasing a patient's survival time compared to if they had not been screened. In this case, the survival time for a patient is increased due to the fact that the patient is known to have cancer for a longer period, and not necessarily due to improvements in care lengthening the patient's life. When making comparisons across time periods and countries, these differences need to be considered before drawing firm conclusions. However, it should be noted that the cure proportion is not susceptible to this bias. Therefore, conclusions can be made about the improvements in the proportion cured without being concerned by the issue of lead time.

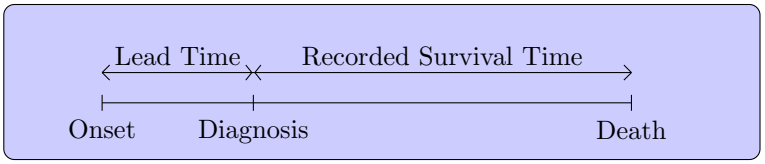

FIGURE 6.5. Illustration of lead time bias for a case of cancer.

6.8.5. Flexible Parametric Modelling of the Cure Proportion

An extension of the flexible parametric models to incorporate the estimate of the cure proportion has recently been undertaken [Andersson et al., 2011]. The restriction of the cubic splines used to model the baseline hazard is used to provide an estimate of the proportion cured by forcing the shape of the curve to be constant beyond the boundary knot.

The approach using the flexible parametric approach is derived as follows. Using the definition for survival in Equation (6.18), and transferring to relative survival:

$$R(t) = \exp(-\exp(\eta)), \tag{6.29}$$

where

$$\eta = s\left(\ln(t)|\gamma, \mathbf{k_0}\right) + \mathbf{x}\boldsymbol{\beta} = \ln(\Lambda(t|z)). \tag{6.30}$$

estimates the log cumulative excess hazard function. The spline function, $s\left(\ln(t)|\gamma, \mathbf{k_0}\right)$ is linear before the first boundary knot, and after the final knot. However, before the first knot, all of the spline variables are zero except for the linear spline variable, $B_1(\ln(t)) = \ln(t)$ (See Section 2.8.1). Andersson *et al.* [2011] propose calculating the splines "backwards" and applying a constraint so that the linear spline variable is 0, forcing the new spline function to just involve a constant beyond the final knot.

Using the constrained "backward" spline function, leads to a relative survival function of the form:

$$R(t) = \pi^{\exp(\sum_{k=2}^{K-1} \gamma_k B_k(\ln(t)))}, \tag{6.31}$$

where $\pi = \exp(-\exp(\gamma_0))$, and γ_0 is the constant term for the constrained "backward" spline function. Looking at Equation (6.27) it is clear that the flexible parametric cure model is a special case of the non-mixture cure model. Using Equation (6.28), it is possible to estimate the survival time of those "bound-to-die" as well as the cure proportion.

The curves fitted using the mixture model in the example in the previous section can be overlaid to those produced by the flexible parametric cure approach (`cure` option of `stpm2` in Stata [StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP]) to show the consistency between the estimates. It has also been shown that using the flexible parametric models for estimating cure gives improvements on the mixture models in certain scenarios [Andersson et al., 2011]. The mixture model often fits poorly to the oldest age groups, and in the case when the relative survival is close to 1. However, it has been shown that in both of these cases, the flexible parametric approach shows improvements [Andersson et al., 2011].

The resulting estimates from the flexible parametric cure model can be compared to the mixture model estimates obtained in the earlier example. Figures 6.6 and 6.7 show that the estimates from the two methods are virtually equivalent for these cases.

However, if we create an older age category for the 75+ age-group, it has been acknowledged that the standard cure methods do not fit well to the data for the oldest ages, even when cure

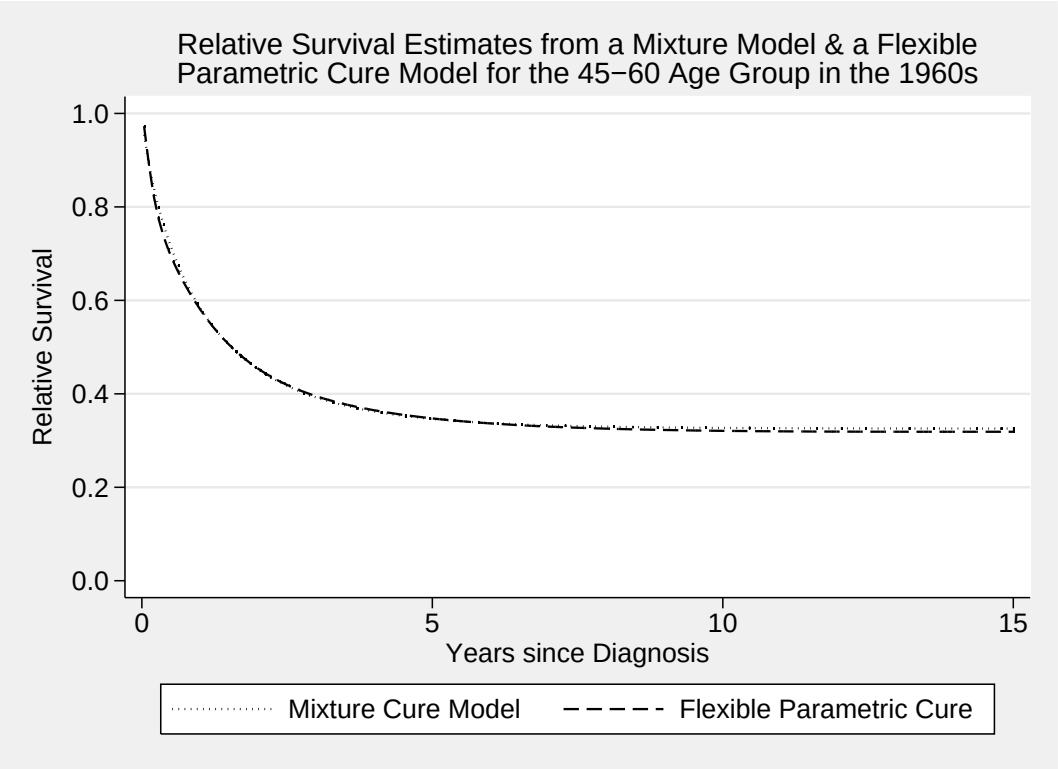

FIGURE 6.6. Cure in the 1960s. Comparing the flexible parametric approach to the mixture model.

appears to be a reasonable assumption [De Angelis et al., 1999; Lambert et al., 2007]. Figures 6.8 and 6.9 show the fitted estimates for the 75+ age-group. The Ederer II estimates are overlaid for comparison. It can be seen that the flexible parametric approach fits better to the lifetable estimates than the mixture model.

6.8.6. Incorporating Cure for Cancer Burden Estimation

The models that are described above assume that the point of cure is reached at time= ∞. This is not a useful estimate of the time to cure when trying to appropriately adjust prevalence estimates for the fact that patients experience cure. Another issue with trying to incorporate cure into the estimation of prevalence is that for certain cancers; such as breast, cure is not actually a reasonable assumption. Women who are diagnosed with a case of breast cancer are at an increased risk of death compared to the general population for many years, and it has been argued that a cure point is not actually reached even after 20 years from the diagnosis [Brenner and Hakulinen, 2004; Woods et al., 2009].

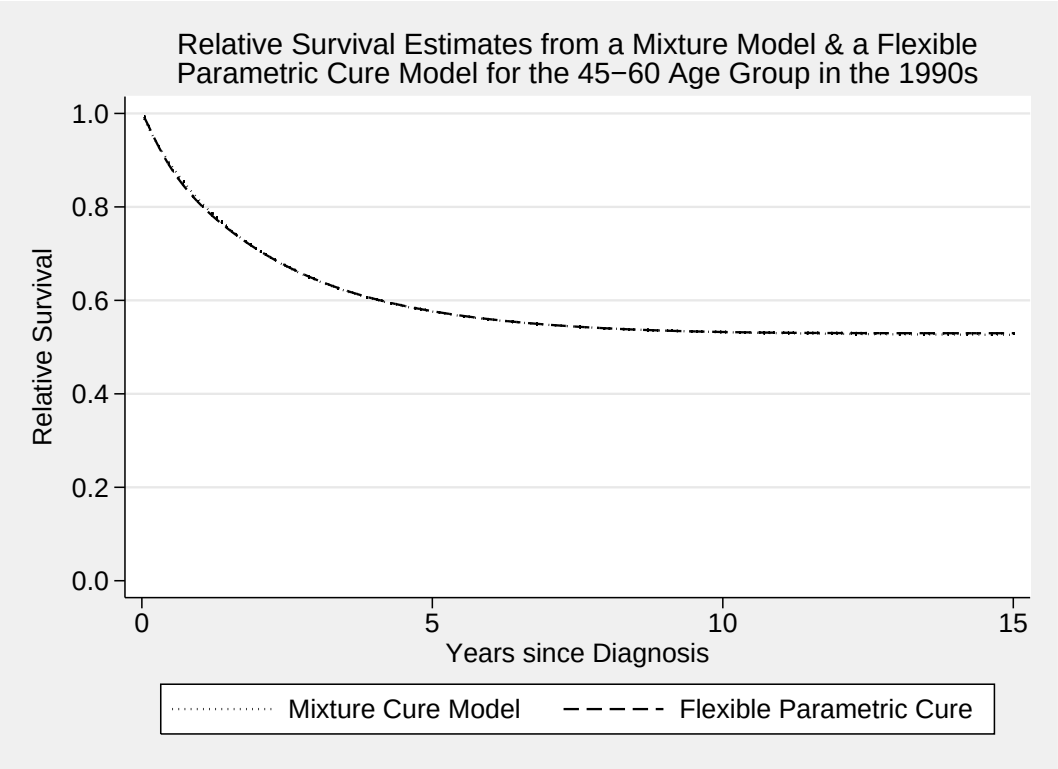

FIGURE 6.7. Cure in the 1990s. Comparing the flexible parametric approach to the mixture model.

In spite of these issues, it is clear that some adjustment for the fact that patients are cured is essential. The prevalence estimates obtained if cure is not accounted for will be over-estimates for the majority of cancers. A patient could still be contributing to the prevalence estimates 50 years after their initial diagnosis if they continue to live on after a cancer diagnosis at a young age. Is this patient likely to still have an active case of cancer in the case of most cancer sites? Is this patient likely to have been a burden due to their cancer diagnosis for all this time? These questions highlight the need for an appropriate adjustment for cure. Partial prevalence estimates do make this adjustment (See Chapter 9). However, the decision to calculate the 5-year or 10-year partial prevalence is somewhat arbitrary without appropriately fitting a cure model to assess the most appropriate cut-off for the time to cure. Can we obtain an appropriate estimate of time-to-cure? It is conceded that the most appropriate method may well be to use external evidence; such as the advice of clinicians, when trying to calculate what would be an appropriate point beyond which cure could be a reasonable assumption. However, the models

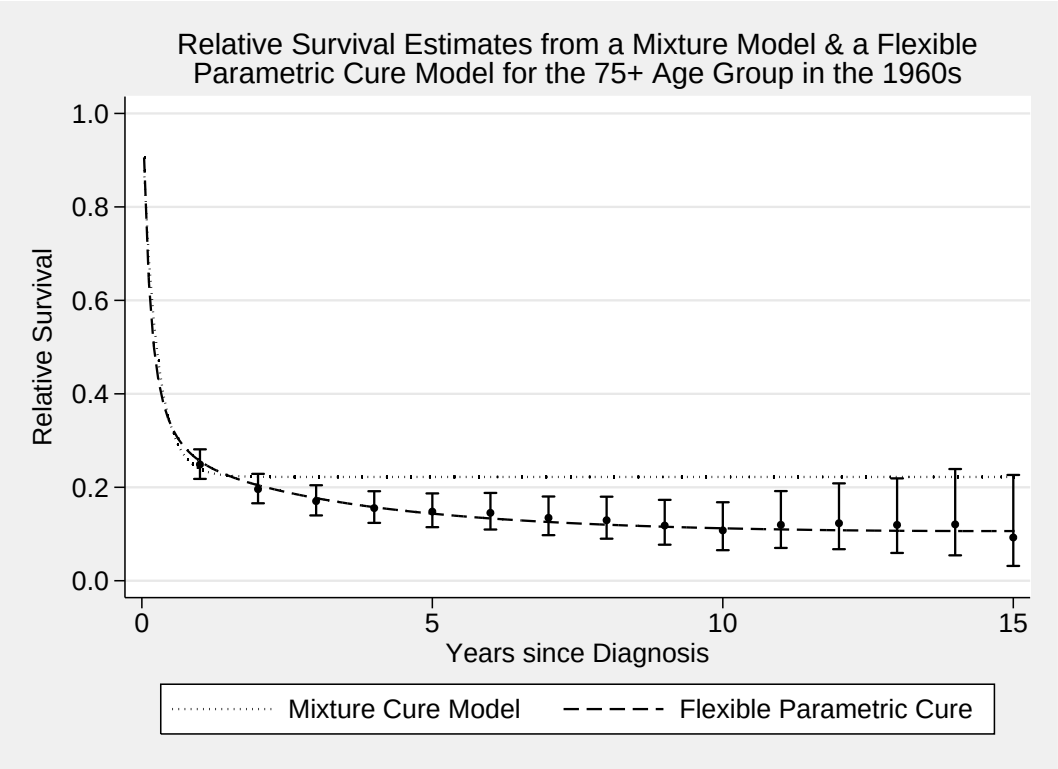

FIGURE 6.8. Cure in the 1960s. Comparing the flexible parametric approach to the mixture model for patients aged 75+.

can be used to assess whether or not cure is a reasonable assumption, and also can be used to visualise the point at which a plateau does appear to have been reached.

6.9. Discussion

In this chapter, the key concepts surrounding quantifying the survival experience of patients in population-based cancer studies have been introduced. In the preceding chapters, the use of population-based cancer data to estimate the rate of incidence of cancer has been explored. Methods to project the estimates of incidence into the future have also been compared in order to build towards a future estimate of cancer burden. The estimation of the survival experience of the patients diagnosed with a new cases of cancer is a vital component of estimating the burden of cancer. In the following chapters, the key concepts of relative survival (Chapter 7), period analysis, modelled period analysis (Chapter 8) and cure (discussed in Chapter 9) will

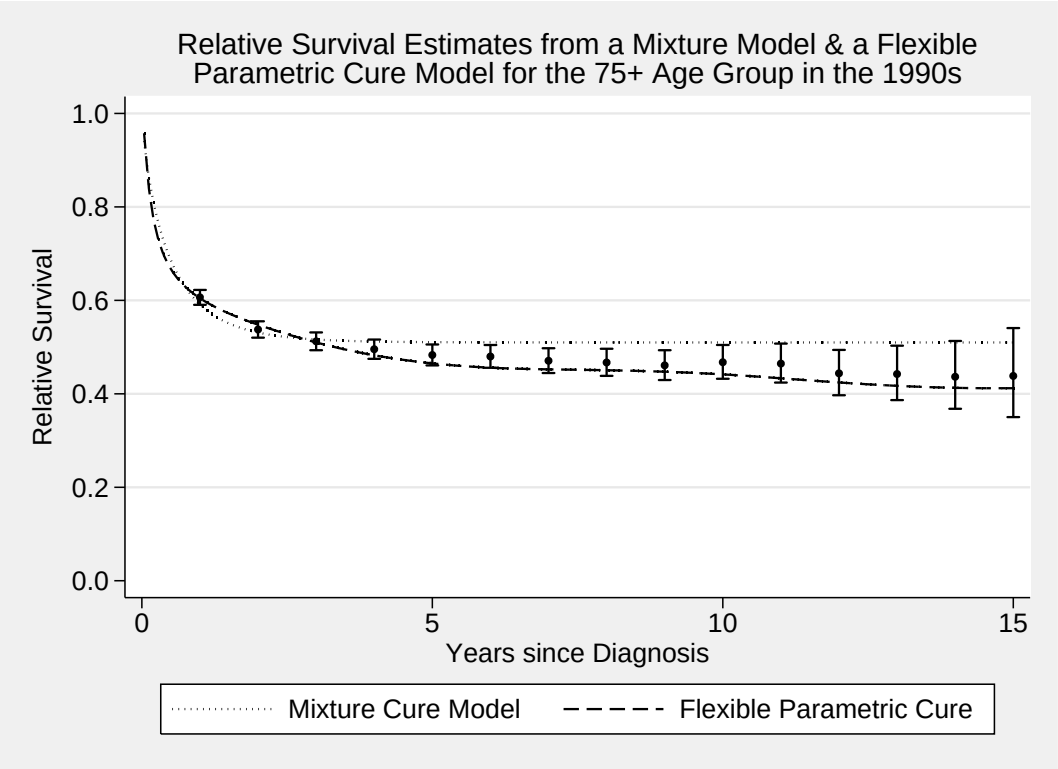

FIGURE 6.9. Cure in the 1990s. Comparing the flexible parametric approach to the mixture model for patients aged 75+.

be explored in detail. An estimate of the survival experience of patients in the future is also necessary to compose an estimate of prevalence and to begin to quantify burden.

The fact that the flexible parametric models can be used to 1) estimate overall, relative or cause-specific survival; 2) can be used for period analysis; and 3) can be used for estimating the cure proportion; means that this modelling framework seems ideal for the purpose of obtaining survival estimates to be used in estimating, and projecting, prevalence.

CHAPTER 7

A Simulation to Compare Methods for Estimating Relative Survival

7.1. Chapter Outline

In this chapter, a full comparison of the approaches used to estimate relative survival is undertaken. Appropriately summarising relative survival is key for cancer registries and the standard methods of estimation, that are used interchangeably, can lead to substantially different estimates. A simulation is performed to try to understand the reasons behind the differences and the importance of taking age into account when reporting relative survival is stressed. This chapter is largely based on a recent publication [Rutherford et al., 2011a]. A copy of the In Press version of this article is given in Appendix IV.

7.2. Introduction

Relative survival can be used as part of a two-stage process of estimating the prevalence of cancer combined with the incidence rate estimates detailed in Chapter 9. However, there are a number of methods that have been suggested for calculating relative survival, and the various methods have been shown to give different estimates. The aim of this chapter is to review and understand the differences between the methods and, through simulation, show the main drivers of these differences. This work is mainly directed at the comparison of estimates of 5 year relative survival that are made available by cancer registries.

Relative survival is the ratio of the observed survival proportion to the expected survival proportion, and tries to give a measure of the probability of surviving a given disease in the absence of other causes. This can be expressed on the hazard scale by considering the excess mortality that is associated with a disease above and beyond what would be expected for a matched patient in the general population. The most common estimates of relative survival are obtained from life-tables using one of three established methods (Ederer I [Ederer et al.,

1961], Ederer II [Ederer and Heise, 1959], or the Hakulinen method [Hakulinen, 1982]). These three methods differ due to the method of calculation for the expected survival. However, with the improvement in computer power, and the ability to analyse relative survival at the individual level, it has been suggested that a modelling approach to relative survival would give more appropriate estimates [Estève et al., 1990] and allow for more complex scenarios to be incorporated.

Relative survival varies by age for the majority of cancer sites. However, it is often of interest for cancer registries to produce a single estimate of relative survival for each cancer site; for example, an estimate of 5 year relative survival. To ensure that only one number is produced for comparisons, these figures are either calculated by pooling all age-groups together, or by using some method of age-standardisation for a set of age(-group)-specific estimates. In this chapter, the estimate that pools all the age-groups together is referred to as the "all-age" estimate. In the literature, this has often been referred to as the crude estimate of relative survival, however, this terminology may cause unnecessary confusion.

There are examples in the literature of pooling all ages, and fitting a grouped modelling approach [Gondos et al., 2009b; Brenner and Hakulinen, 2008; Lambert et al., 2005]. There are also examples in which life-table estimates are used that pool all age-groups together [Talbäck et al., 2004]. The majority of the recent examples that pool age relate to the method proposed by Brenner for obtaining up-to-date estimates of relative survival using period modelling [Brenner and Hakulinen, 2008; Pulte et al., 2010]. This method uses a grouped modelling approach whilst usually pooling all ages together. Further details of this approach will be given in the following chapter; Chapter 8.

In this chapter, the differences that can occur by using the various methods are highlighted through a motivating example for cancer of the thyroid gland from Finland. A simulation study is performed in order to investigate the main drivers behind these observed differences. This chapter emphasises the need to calculate prevalence in appropriate age categorisations, as the estimates of relative survival for all ages combined do not give an estimate of net survival, which is required to ensure the prevalence calculations are correct.

7.3. Methods

Relative survival methods are used to try to obtain an estimate of net survival; that is, the probability of surviving the disease of interest in the absence of death from other causes. This concept was introduced in the previous chapter, in Section 6.5. Net survival is a hypothetical measure, but is useful for comparisons between groups in that it is adjusted for the fact that different populations may have different levels of background risk of death. The expected (or background) mortality is normally obtained from nationwide or regional population mortality figures, which are obtained in yearly intervals for age and calendar time, whilst also being estimated separately for each sex. Relative survival as a function of time ($R(t)$) is defined as:

$$R(t) = \frac{S(t)}{S^*(t)}, \tag{7.1}$$

where $S^*(t)$ is the background survival in the population, and $S(t)$ is the observed survival for the cancer patients. This can also be written on the hazard scale, with the excess hazard (mortality), $\lambda(t)$, defined by:

$$\lambda(t) = h(t) - h^*(t), \tag{7.2}$$

where $h(t)$ is the observed hazard (mortality) amongst the cancer patients, and $h^*(t)$ is the background hazard (mortality) for a relevant comparative population. This shows how the total mortality, $h(t)$, is split into two components; the background mortality due to other causes, $h^*(t)$, and the excess mortality associated with the disease of interest, $\lambda(t)$.

7.3.1. Life-table Approaches

The three life-table approaches (Ederer I [Ederer et al., 1961], Ederer II [Ederer and Heise, 1959], and the Hakulinen method [Hakulinen, 1982]) differ purely in their estimation of $S^*(t)$ when calculating relative survival using equation (7.1). The time-scale is usually split into yearly intervals, with separate estimates made for each time interval.

The Ederer I [Ederer et al., 1961] approach is the simplest method of estimating the expected survival proportion. It makes the assumption that the time at which a cancer patient dies or is censored has no effect on the expected (background) survival. The Ederer I approach does

not allow for the fact that the patients have heterogeneous follow-up, which can lead to biased estimates of relative survival [Hakulinen, 1982].

For the cumulative expected survival from the beginning of follow-up until the end of the i^{th} interval, cp_i^*, under the Ederer I approach the formula is given by:

$$cp_{\text{E1}i}^* = \frac{1}{l_1} \sum_{k}^{l_1} \left(\prod_{j=1}^{i} p_j^*(k) \right), \tag{7.3}$$

where l_1 is the total number of patients alive at the start of follow-up, and $p_j^*(k)$ is the expected survival for the j^{th} interval according to population life-tables for a patient similar to the k^{th} patient with respect to the variables in the life-table; often age, sex and calendar year. The product is calculated for all i intervals for each of the l_1 patients, even if the patient dies or is censored in the first interval.

The Ederer II [Ederer and Heise, 1959] approach controls for heterogenous observed follow-up times by appropriately accounting for when the matched individuals should be at risk. However, the expected survival proportion is dependent on the observed mortality [Hakulinen, 1982], which leads to potentially biased estimates of relative survival. This often has little effect in practice, and it has recently been argued that the Ederer II approach should be the preferred life-table approach for estimating relative survival [Hakulinen et al., 2011].

For the cumulative expected survival from the beginning of follow-up until the end of the i^{th} interval, cp_i^*, under the Ederer II approach the formula is given by:

$$cp_{\text{E2}i}^* = \prod_{j=1}^{i} \left(\frac{1}{l_j} \sum_{k}^{l_j} p_j^*(k) \right), \tag{7.4}$$

where l_j is the total number of patients alive at the start of j^{th} interval, and $p_j^*(k)$ is the expected survival for the j^{th} interval according to population life-tables for a patient similar to the k^{th} patient with respect to the variables in the life-table; often age, sex and calendar year. That is, the estimate of expected survival in each interval is calculated only for patients alive at the start of each interval.

Table 7.1 gives an example of a life-table for a relative survival calculation for the Ederer II estimate. The table shows 5 yearly intervals in terms of the observed and expected number of deaths in each interval. The effective number at risk column gives the value that is used

Start	End	Alive at Start	Obs. Deaths	Exp. Deaths	Censored	Effective at Risk	p^* Expected	p Observed	r Relative	cr Cum. Rel. (Ed. II)
0	1	6274	780	65.9	2	6273	0.9842	0.8757	0.8897	0.8897
1	2	5492	158	55.1	306	5359	0.9892	0.9704	0.9810	0.8728
2	3	5028	99	50.2	338	4859	0.9895	0.9796	0.9900	0.8641
3	4	4591	80	46.7	303	4439.5	0.9894	0.9820	0.9925	0.8576
4	5	4208	61	43.1	255	4080.5	0.9893	0.9851	0.9957	0.8530

TABLE 7.1. Example of Life-Table

for the denominator in the survival calculations; this column is calculated under the acturial assumption for the censoring pattern in the intervals. The observed survival column can be obtained by calculating $\left(1 - \left(\frac{\text{Observed Deaths}}{\text{Effective at Risk}}\right)\right)$. The interval specific relative survival, r, can be calculated by $\frac{p}{p^*}$. The cumulative relative survival, cr, is calculated by the cumulative multiplication of the r column.

The Hakulinen method [Hakulinen, 1982] for estimating relative survival adjusts for potentially heterogenous follow-up times and, in this case, expected survival is independent of the observed mortality of patients. This is achieved by estimating a biased estimate of the expected survival proportion to account for the fact that the observed survival proportion is equally biased.

When pooling all ages, it has been shown that the Hakulinen (and Ederer I) estimates tend towards the survival proportion of the youngest patients as follow-up time increases [Hakulinen, 1977]. This can lead to an increasing slope for the all-age relative survival curve that is not seen in any of the age-group-specific curves. This is demonstrated in the following example in Figure 7.2.

A recent approach that has been suggested claims to directly estimate net survival when the appropriate life-table is used [Perme et al., 2011]. This approach has been proposed as a continuous time approach but has been recently added to a life-table estimation command in the statistical software package, Stata [StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP]. Unfortunately, this approach was suggested after the simulation was carried out. However, the Pohar-Perme approach has been included for the real example. It should be noted that the estimates given in the real example were made using the experimental code in the `strs` software in Stata, and further testing is needed of this functionality of the command.

The Pohar-Perme estimator of relative survival is calculated by applying weights to the Ederer II approach. Define individual patient weights for each subject $w_{ij} = \frac{1}{S^*_{ij}}$, where i indicates each subject, and j the follow-up interval. The Pohar-Perme approach uses these weights for the overall and expected survival in order to inflate the number of patients in the risk-set for each interval to account for the fact that patients are dying from causes other than cancer. On the basis, that the estimate required should be in the hypothetical world where cancer is the only cause of death, it seems an intuitive solution to estimating the net survival quantity.

7.3.2. Poisson Modelling

A Poisson modelling approach can also be used to provide an estimate of relative survival from registry data [Dickman et al., 2004]. The Poisson model can be applied to either grouped-level data or individual-level data. The grouped-modelled approach is applied to life-table level data and is similar in theory to the life-table approach set out by Reeves *et al.* [1999].

Rearranging Equation 7.2 for the excess hazards:

$$h(t) = h^*(t) + \lambda(t). \tag{7.5}$$

Assuming that the number of deaths for observation j follows a Poisson distribution, $d_j \sim$ Poisson(μ_j), and using $\mathbf{x}\boldsymbol{\beta}$ to model the log excess hazard rate, Equation (7.5) can be written as:

$$\frac{\mu_j}{y_j} = \frac{d^*_j}{y_j} + \exp(\mathbf{x}\boldsymbol{\beta}). \tag{7.6}$$

which can be rearranged as:

$$\ln(\mu_j - d^*_j) = \ln(y_j) + \mathbf{x}\boldsymbol{\beta}. \tag{7.7}$$

This implies a GLM with outcome, d_j, a Poisson error structure, and a non-standard link function $\ln(\mu_j - d^*_j)$ and an offset of the the log person-time, $\ln(y_j)$ [Dickman et al., 2004]. The subscript j can either be used to describe an individual subject, or alternatively, the j subscript can refer to grouped level data, collapsed over covariate patterns. The two approaches lead to

different estimates in the case of relative survival estimation and the approaches are constrasted below.

In the explanation of the difference between the individual and grouped model that follows, a single time-interval is considered for the estimation of the excess mortality. However, it is generalisable to multiple time-intervals. For the all-age grouped-based Poisson model, a single excess mortality rate is estimated, λ, within each follow-up interval, such that:

$$h_i = \overline{h^*} + \lambda. \tag{7.8}$$

That is, each individual is assumed to have a background mortality that is equal to the average background mortality. This seems to be a strong assumption in the light of the known differences in terms of expected mortality for individuals of different ages. This approach can be applied separately to a pre-defined set of age-groups, or the estimates can be modelled by assuming proportional excess hazards.

In contrast, for the all-age individual-based Poisson model, a single excess mortality rate is estimated, λ, within each follow-up interval, such that:

$$h_i = h_i^* + \lambda, \tag{7.9}$$

Since h^* now has a subscript, i, this allows each individual to have their own, individual background hazard when calculating the value of λ. This seems to be a more appropriate assumption than the one made by equation (7.8). In fact, if the average hazard for a follow-up interval is calculated and applied to each individual then we can replicate the results obtained from the grouped modelling approach when using individual data. As with the grouped approach, it is possible to fit this model for separate age-groups, which would be equivalent to a full interaction with age-group, or use the estimates from separate age-groups in a proportional excess hazards model. The individual approach proposed by Estève *et al.* has been shown to be equivalent to a Poisson modelling approach using individual level data [Dickman et al., 2004].

The Poisson modelling approaches require splitting of the time-scale into intervals in order to get estimates of relative survival. The standard approach is to split the timescale into yearly intervals, as is commonly done for the standard life-table estimates. It is assumed that the excess hazard rate is constant within each interval, which may be inappropriate, particularly

within the first year of follow-up. Therefore, another option is to finely split the time-scale to get a more appropriate estimate of relative survival. On the basis that, for most cancer sites, the largest change in the hazard occurs early on in follow-up, data (grouped and individual) that are split finely in the first year of follow-up and in yearly intervals thereafter are also used. The finely split models will split the first year of follow-up into 12 equally-spaced intervals.

7.3.3. Hakulinen-Tenkanen Model

A further modelling approach has been suggested that uses the grouped-level data to estimate relative survival [Hakulinen and Tenkanen, 1987]. The Hakulinen-Tenkanen model is a modelling extension to the Ederer II approach. This model fits into the framework of generalised linear models by assuming the number of patients surviving the interval follows a binomial distribution with denominator the effective number at risk and using a complementary log-log link [Dickman et al., 2004]. This modelling framework has the advantage of not assuming that the excess hazard is constant for each of the intervals that define the group-level data, which contrasts to the Poisson modelling approach.

For a single stratum of a life-table, let $l_i' - d_i$ be the number of patients surviving the interval, where l_i' is the effective number at risk accounting for censoring using the actuarial assumption. The model can be defined as a GLM with outcome $l_i' - d_i$, a binomial error structure, and a complementary log-log link function:

$$\ln\left\{-\ln\frac{p_i}{p_i^*}\right\} = \mathbf{x}\boldsymbol{\beta}. \tag{7.10}$$

7.3.4. Flexible Parametric Models

A further individual modelling approach is the flexible parametric approach proposed by Royston and Parmar [Royston and Parmar, 2002; Nelson et al., 2007]. The mathematical details of the flexible parametric approach to relative survival are given in Section 6.5.4. This approach has the advantage of treating time continuously meaning that an arbitrary splitting of the time-scale is not required. The approach uses restricted cubic splines to describe the shape of the baseline hazard [Nelson et al., 2007]. Using splines means that a choice upon the number of knots is required. For the analyses carried out in this chapter, 5 degrees of freedom (4 internal knots) will be used for the flexible parametric models. The issue of the choice of knots for these models is not a major concern provided that a sufficient number of knots are used [Lambert

and Royston, 2009]. This modelling approach is very flexible and is particularly useful for the incorporation of time-dependent effects. It also has the distinct advantage of not requiring the time-scale to be split as time is treated continuously in this case.

7.3.5. Age Standardisation

To make formal comparisons between an all-age and age-standardised estimate, internal traditional age-standardisation was carried out for each of the analyses described. Traditional age-standardisation weights the age-group-specific estimates according to the initial age structure of the cohort. Both the Poisson approaches and the flexible parametric approach will assume proportional excess hazards when modelling the effect of age throughout the analyses carried out.

Traditional age-standardisation of relative survival can be defined mathematically as follows. Let w_a be the proportion of patients in age-group a at the start of follow-up. With A age-groups, the age-standardised relative survival estimate at time-point t, $R_s(t)$, can be given as:

$$R_s(t) = \frac{\sum_{a=1}^{A} w_a R_a(t)}{\sum_{a=1}^{A} w_a}, \tag{7.11}$$

where $\sum_{a=1}^{A} w_a = 1$, and $R_a(t)$ is the age-group specific relative survival estimate at time t in age-group a.

Other age-standardisation methods are available [Brenner and Hakulinen, 2003; Brenner et al., 2004a]. The more recent method suggested by Brenner [Brenner et al., 2004a] would provide exactly the same estimates as the all-age estimates that are contained in this chapter, provided that internal age-standardisation was carried out. On this basis, the results of this standardisation method are not explicitly given as part of the results. The earlier method suggested by Brenner [Brenner and Hakulinen, 2003] has weights that alter throughout follow-up to account for the fact that the age distribution is not constant with time. This can lead to the estimate of relative survival converging to the estimate for the youngest age-group for longer-term follow-up [Pokhrel and Hakulinen, 2008].

7.4. Motivating Example

An analysis was performed on patients diagnosed with cancer of the thyroid gland from data made available by the Finnish Cancer Registry. A comparison of the estimates given from

both the life-table and modelled approaches was made in terms of five year relative survival for patients diagnosed between 1985 and the end of 2004 (with follow-up until the end of 2005). The five-year relative survival estimates are presented in Table 7.2. The all-age grouped Poisson estimate (84.60%) is substantially lower than the estimates obtained from the individual grouped Poisson (89.78%) and the flexible parametric (89.31%) approaches. Further differences can be observed for the all-age estimates available from the life-table methods. Using the fine-splitting for the first year of follow-up for the Poisson models leads to a difference of about 0.4 percentage units for both the grouped and individual approaches.

Although the age-standardised estimates are more similar for each of the methods, there is still some difference. This could potentially be due to the fact that the proportional excess hazards assumption may not hold, which would affect each of the estimates that were obtained from a model. The life-table approaches would not be affected by this issue as they perform a separate analysis for each of the age-groups. It is, of course, possible to relax the proportional excess hazards assumption for each of the modelling approaches. However, relaxing the proportional hazards assumption made little difference to the age-standardised estimates. This is unsurprising considering the results of Table 7.2. The Hakulinen-Tenkanen standardised estimate (84.72%) is very similar to the Ederer II age-standardised estimate (84.66%). These approaches give equivalent estimates for the all-age approach and the only difference for the standardised estimates is that the Hakulinen-Tenkanen approach makes the proportional hazards assumption.

The Pohar-Perme all-age estimate (84.06%) gives the true net survival estimate. The fact that some difference is seen between this estimate and the age-standardised estimates from the other methods indicates that there is persistent heterogeneity in the age-groupings that have been used. Looking further at the resulting age-specific group estimates highlights that the difference occurs due to the open-ended age-group (75+). Further splitting this age-group into more homogeneous sub-groups would lead to the age-standardised approaches giving more consistent estimates, and also lead to them giving a similar estimate obtained from the Pohar-Perme approach. Thyroid cancer is an extreme example; the age-standardised estimates will usually estimate the net survival well.

Method	All-Age 5 Year RS	Trad. Stand.
Ederer I	87.39	85.01
Ederer II	85.39	84.66
Hakulinen	87.55	85.10
Grouped Poisson	84.60	83.81
Ind. Poisson	89.78	84.53
Hakulinen-Tenkanen	85.39	84.72
Grouped Poisson (Fine)	84.96	84.61
Ind. Poisson (Fine)	89.35	85.27
Flexible Parametric Model	89.31	85.30
Pohar-Perme	84.06	-

TABLE 7.2. All-Age 5 Year Relative Survival; Cancer of the Thyroid Gland in Finland for Patients Diagnosed between 1985 and 2004 (with follow-up until the end of 2005).

Further differences of the various methods can be observed if a longer follow-up interval is used. The results for up to 15 years of follow-up are presented graphically in Figure 7.1. The Ederer I and Hakulinen estimates increase after around 4 years of follow-up until the end of the follow-up. However, this increase is not observed at the same point in any of the age-specific curves (see Figure 7.2). This is due to the fact that as follow-up increases the Ederer I and Hakulinen estimates tend towards the observed effect for the youngest patients [Hakulinen, 1977] because these patients receive greater weight in these approaches as they are more likely to be long-term survivors. This problem is not evident in the model-based approaches, and is also not an issue for the Ederer II estimates [Hakulinen et al., 2011]. The graphical representation highlights the strong similarity in the estimates for the two individual-based approaches; and the further similarity between the estimates of the grouped-based Poisson model and the Ederer II approach.

When the individual-based Poisson approach is fine-split for the first year of follow-up, the 5 year relative survival is very similar to the flexible parametric approach. The flexible parametric approach does not require an arbitrary choice of the splitting of the timescale. However, it should be conceded that there is a requirement to select the number of knots that are used to model the baseline hazard function using the spline variables. This, however, has been shown to make little difference in terms of the estimated relative survival providing that a sufficient number of knots are used [Lambert and Royston, 2009]. In this instance, the all-age

estimate of relative survival using 4 degrees of freedom is 89.33%, and for 6 degrees of freedom is 89.35%.

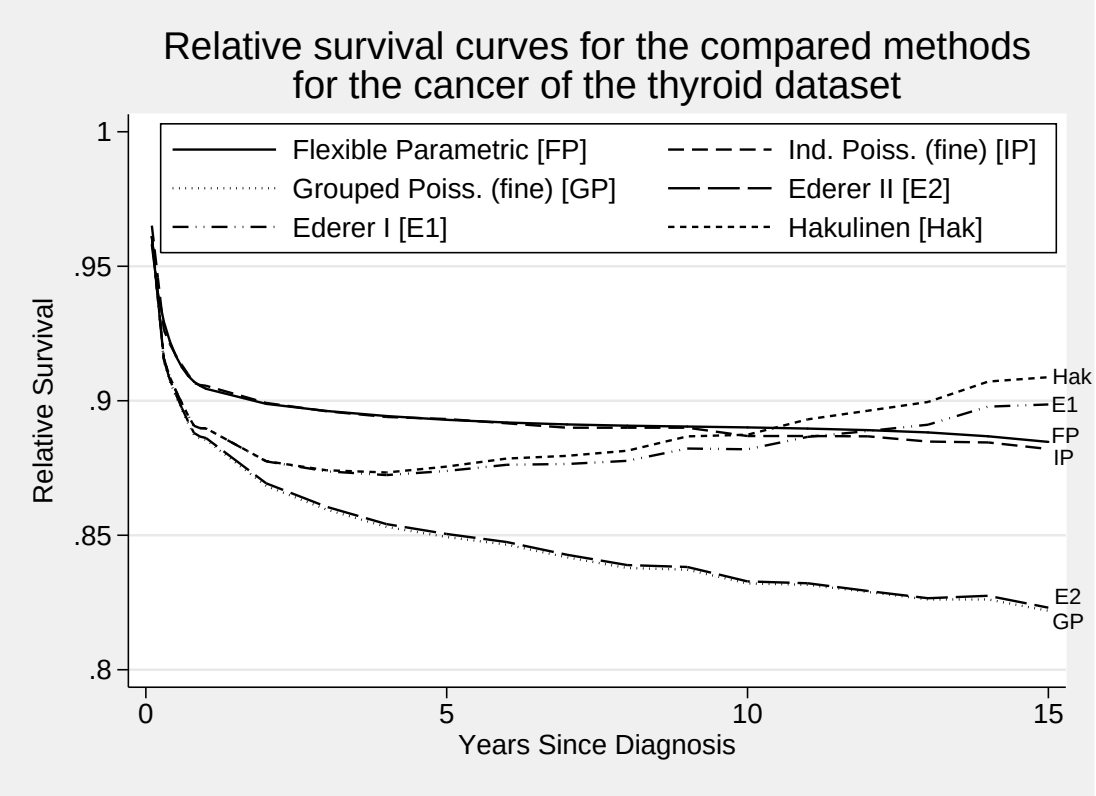

FIGURE 7.1. Relative Survival Curves for Cancer of the Thyroid Gland in Finland for Patients Diagnosed between 1985 and 2004.

The differences between the grouped-based and individual-based relative survival estimates at 15 years of follow-up are in the region of 6 percentage units. Each of the methods presented in the graphs are usually assumed to be various attempts at estimating the same quantity. It can be seen from this example that, under some circumstances, substantially different estimates can be obtained. Understanding the assumptions, and the conditions that are likely to lead to disparate estimates is vital to selecting the most appropriate method of analysis.

7.5. Simulation Study

A simulation study was carried out to attempt to uncover the driving forces for the differences between the life-table estimates, and the various individual and grouped model-based estimates of relative survival. Eight simulation scenarios were defined to compare the effect of

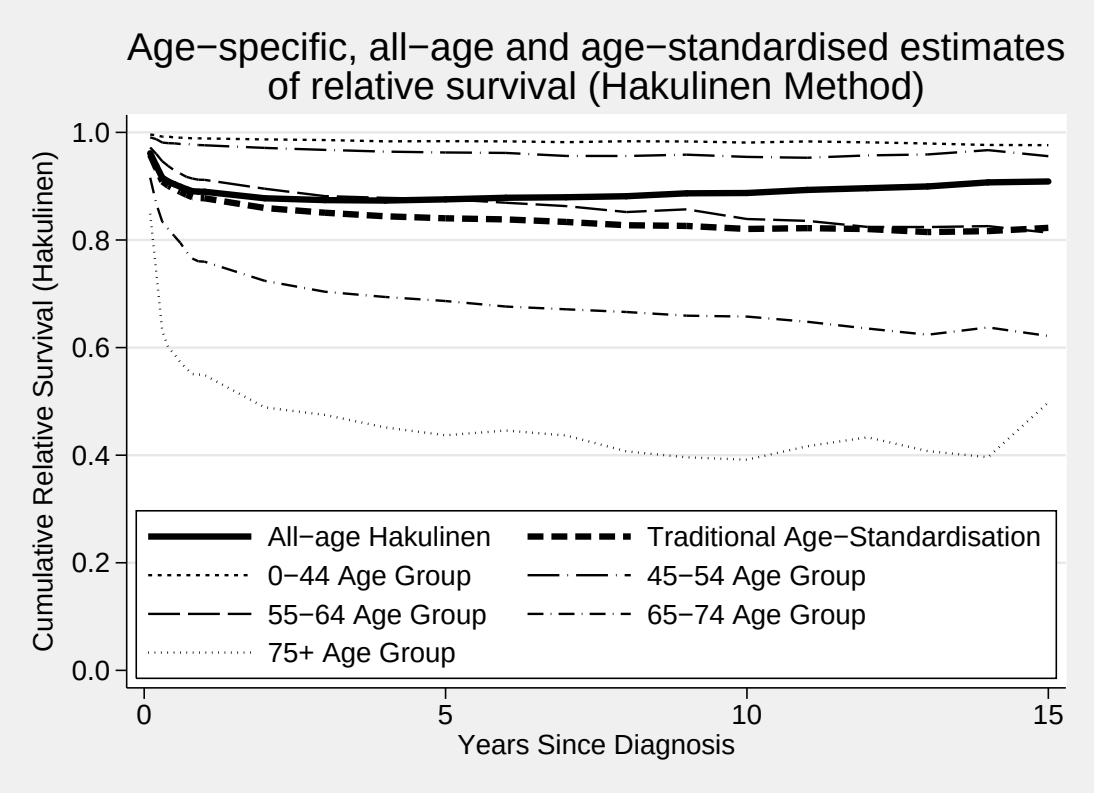

FIGURE 7.2. Relative Survival Curves for Cancer of the Thyroid Gland in Finland for Diagnoses between 1985 and 2004.

a) the shape of the relative survival curve (and consequent value of 5 year relative survival); b) the effect of the age distribution of the patients; and c) to evaluate the effect of allowing the relative survival to be dependent on the age structure (that is, defining excess hazard ratios for a pre-defined set of age groups).

In order to simulate relative survival, each individual was simulated a time to death due to cancer, and a time to death due to other causes; the minimum value from these two was then taken as the value for the time to death for that individual. The simulation was carried out as follows:

(1) Times of death due to cancer were generated from a Weibull distribution according to the method set out by Bender *et al.* [2005] Different shapes for survival curves were generated by altering the parameter values of the Weibull distribution, referred to as the scale and shape parameters, commonly denoted λ and γ respectively.

(2) Time to death due to other causes was calculated by using a Finnish population mortality file and using an Exponential distribution for each attained age during follow-up.

(3) Overall time to death was calculated by taking the minimum of the cancer-specific time to death, and the expected (background) time to death.

(4) Two age distributions were simulated from normal distributions with a given mean, and standard deviation.

(5) The effect of age was simulated by using pre-selected excess hazard ratios for the defined age-groups ($\leqslant$44, 45-54, 55-64, 65-74, $\geqslant$75) with the central age-group as the reference.

(6) Each method for calculating relative survival was applied to each of the simulated datasets, calculating all-age estimates and using traditional age-standardisation for each method.

Table 7.3 outlines the simulation strategy that was used to investigate each of the scenarios. In total, 8 scenarios were chosen, ensuring that each of the variants were appropriately distributed between them. Those scenarios with a "Yes" for the "Age Effect" column in Table 7.3 were those that had the hazard ratios for the effect of age (0.6, 0.7, 1, 1.7, 3.5; respectively) as part of the simulated data. Those with a "No" for the age effect were assumed to have an equivalent relative survival curve for each of the five age groups ($\leqslant$44, 45-54, 55-64, 65-74, $\geqslant$75). The "Mean Age" column in Table 7.3 indicates the mean of the normal distribution for age, each had a standard deviation of 13. Finally, the "RS (Relative Survival)" column of Table 7.3 refers to the shape of the Weibull distribution that was selected for the simulated relative survival times. The values indicate the λ used for the Weibull distribution. All scenarios were simulated with γ=0.5. Assuming that there is no effect of age, the age-standardised 5-year relative survival estimate for the low relative survival group (λ=0.9) is 13.4%, whereas the equivalent value for the high survival group (λ=0.2) is 63.9%.

Figure 7.3 shows the shape of the survival distribution that is used for each of the scenarios. The top graph shows the resulting curves when there is assumed to be no effect of age. These graphs show the low, and high relative survival curves that are used for scenarios 1-4. The bottom two graphs indicate the survival associated with the 5 age-groups under the two different

Scenario	Age Effect	Mean Age (years)	RS shape (λ for Weibull Dist.)
1	No	60	0.2
2	No	60	0.9
3	No	70	0.2
4	No	70	0.9
5	Yes	60	0.2
6	Yes	60	0.9
7	Yes	70	0.2
8	Yes	70	0.9

TABLE 7.3. Simulation Strategy

Weibull distributions. The left graph of the two shows the high survival curves (λ=0.9), whereas the right graph relates to the low relative survival curves (λ=0.2).

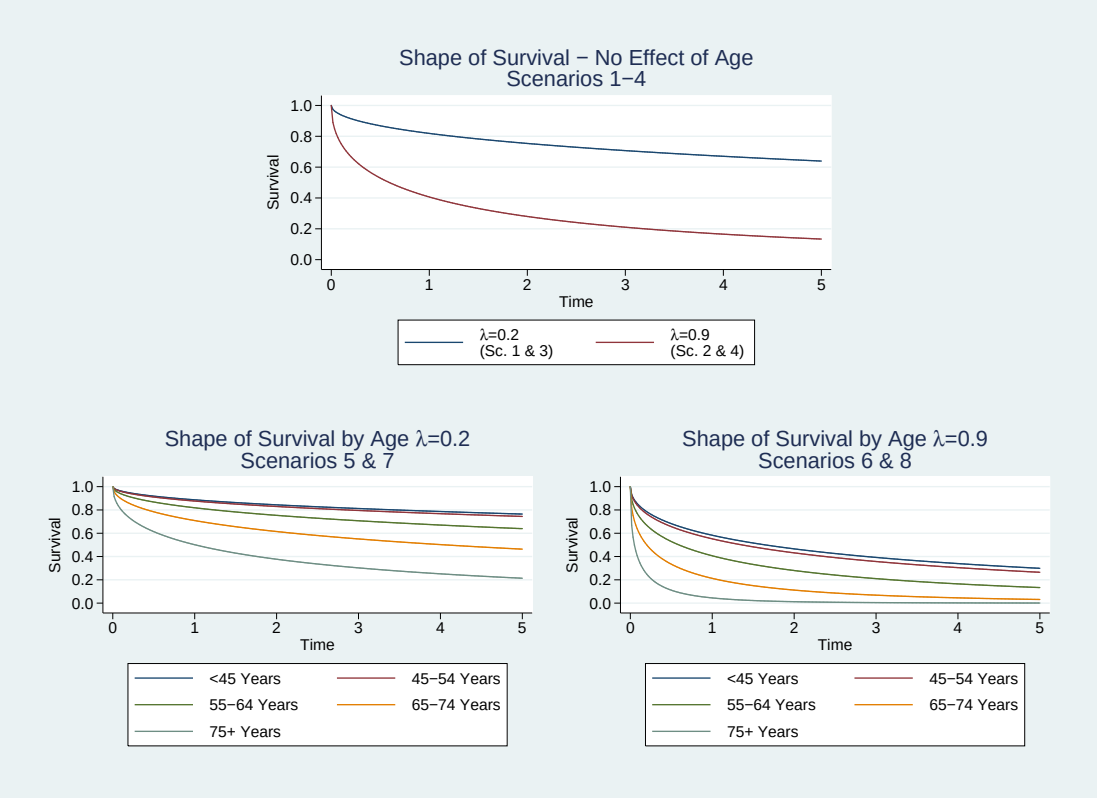

FIGURE 7.3. Shape of the survival curves according to the selected Weibull distributions for the Scenarios.

For each strategy, 500 simulations were run and the 5 year relative survival estimates were stored. A sample size of 10,000 was used for each of the simulations. The results are summarised by the mean of these 500 simulations, as well as by scatter plots of each of the points so the

spread of the estimates can be fully appreciated. The aim of the simulation was to investigate the bias in the estimates, motivated by the results of the real example using thyroid cancer data.

7.6. Simulation Results

The results of the various simulation strategies are reported in Table 7.4. The top half of the table contains the four scenarios where the effect of age was assumed to be the same across all of the age categorisations. The bottom half of Table 7.4 contains the scenarios where age has a proportional effect within the set age categorisations.

Scenarios 1-4 contain the results of the simulations that assumed that there was no effect of age on excess mortality, that is, that patients of all ages were assumed to have the same shape for the simulated relative survival curve. The results highlight that, under these conditions, there is relatively little difference between the estimates produced by any of the life-table or modelling methods. This highlights that the effect of age is a significant determinant of the observed differences between the methods. The conditions investigated in Scenarios 1-4; where age does not have an effect on the excess mortality, are very rare in practice.

The only substantial differences between any of the estimates from the various methods in Scenarios 1-4 are as a result of the splitting of the timescale for the Poisson modelling approaches. The estimates of the fine-split Poisson models (only finely-split for the first year of follow-up), are more in-line with the other estimates than those that were fitted to the yearly split data. This is particularly apparent for the two scenarios where a lower relative survival was used (Scenarios 2 and 4). This is simply due to the fact that there is a large decrease in relative survival in the first year due to the Weibull distribution that was selected in this case. The estimated hazard that is used to calculate the relative survival estimates for the Poisson estimates is changing rapidly in the first year of follow-up (see Figure 7.3), and this is not accounted for in the yearly split methods. The life-table estimates are less affected by finely splitting the time intervals and so only the crudely split estimates are reported. For example, in Scenario 4, the all-age grouped Poisson estimate of 5 year relative survival is 11.259%, whereas the all-age grouped Poisson (Fine) estimate of 5 year relative survival is 13.083%. A similar magnitude is observed for the difference between the finely split individual Poisson method and the yearly split individual method. This difference of nearly 2 percentage

units is considerable. It can also be seen that the fine-splitting issue carries over into the age-standardised estimates. The situation of a rapidly changing excess hazard within the first year of follow-up is mirrored in a number of cancer sites in practice and so this issue is one which may need further attention when Poisson modelling approaches are used. It should be noted that the flexible parametric approach has the advantage of treating time continuously meaning that a decision on the splitting of the timescale is not required. It can be seen that the all-age flexible parametric estimate (13.343%) and the all-age fine-split individual Poisson approach (13.131%) give estimates that are in closer agreement than the yearly split estimates.

	Description	**Mean Age Low**				**Mean Age High**			
		High RS		**Low RS**		**High RS**		**Low RS**	
	Scenario	**1**		**2**		**3**		**4**	
	Method	All-Age	Trad.	All-Age	Trad.	All-Age	Trad.	All-Age	Trad.
No Effect of Age	Ederer I	63.948	63.950	13.379	13.377	63.961	63.964	13.376	13.381
	Ederer II	63.951	63.950	13.379	13.377	63.961	63.961	13.377	13.380
	Hakulinen	63.948	63.950	13.379	13.377	63.961	63.964	13.376	13.381
	Grouped Poisson	63.454	63.453	11.338	11.338	63.396	63.382	11.274	11.259
	Ind. Poisson	63.464	63.458	11.341	11.339	63.431	63.401	11.281	11.266
	Hakulinen-Tenkanen	63.951	63.958	13.379	13.380	63.961	63.976	13.377	13.377
	Grouped Poisson (Fine)	63.816	63.869	13.116	13.160	63.708	63.767	13.083	13.124
	Ind. Poisson (Fine)	63.886	63.877	13.163	13.161	63.823	63.786	13.131	13.127
	Flexible Parametric Model	63.952	63.958	13.335	13.355	63.965	63.971	13.343	13.345
	Truth	63.941	63.941	13.366	13.366	63.94	63.941	13.366	13.366
	Scenario	**5**		**6**		**7**		**8**	
	Method	All-Age	Trad.	All-Age	Trad.	All-Age	Trad.	All-Age	Trad.
Differential Effect of Age	Ederer I	60.281	58.389	14.851	14.385	48.987	44.991	8.534	7.291
	Ederer II	59.128	58.388	14.230	14.385	46.646	44.993	7.573	7.291
	Hakulinen	60.281	58.389	14.851	14.385	48.987	44.991	8.534	7.291
	Grouped Poisson	58.129	57.610	11.114	12.510	44.719	43.665	4.193	6.213
	Ind. Poisson	60.085	57.615	11.345	12.888	47.714	43.679	4.393	6.281
	Hakulinen-Tenkanen	59.128	58.388	14.230	13.967	46.646	44.992	7.573	7.214
	Grouped Poisson (Fine)	58.753	58.298	13.731	13.829	45.893	44.780	6.956	7.118
	Ind. Poisson (Fine)	60.664	58.301	13.991	14.243	48.847	44.805	7.233	7.194
	Flexible Parametric Model	60.755	58.385	14.275	13.972	49.053	44.980	7.570	7.171
	Truth	58.339	58.339	13.992	13.992	44.952	44.952	7.216	7.216

TABLE 7.4. Estimates of 5 Year Relative Survival from the various approaches; Results for all Simulation Scenarios (1 - 8).

Scenarios 5-8 were those that assumed the proportional effect of age using pre-specifed hazard ratios. Scenario 5 assumed to have a specific age distribution; a normal distribution with a mean of 60, and a standard deviation of 13 (See Table 7.3 for Scenario descriptions). The results of this scenario are similar to those seen in the previous section for the thyroid example dataset. However, in this case the Ederer I and Hakulinen estimates are entirely equivalent. This is due to the fact that the simulation was carried out on a complete dataset; that is, there was no censoring to account for other than that defined by the end of follow-up at 5 years.

The all-age Ederer II (59.128%) and all-age grouped Poisson (Fine) estimates (58.753%) are again similar. The all-age Hakulinen (and Ederer I) estimate (60.281%) is slightly higher than this estimate and it is expected that the difference would become larger if follow-up time was increased.

There is a large difference between the all-age grouped Poisson estimates, and the all-age individual Poisson estimates for Scenarios 5-8. This difference is reduced when the estimates are age-standardised. Essentially, for the all-age estimate, an individual background hazard is included for each person in the individual model, whereas in the grouped model the average of the background hazard for all the patients in that follow-up interval is applied to everyone. The grouped model makes an assumption that only holds when the ratio of observed to expected is similar to the ratio of the averages of those values. This is what occurs in Scenarios 1-4 as it was forced as part of the simulation, and this is also what happens when the age(-group)-specific analyses are performed in order to calculate the standardised estimates as the variation in the expected survival is reduced.

The second column for each scenario in Table 7.4 contains the results of traditional age-standardisation. Each of the methods give comparable results. The only differences observed in this column appear to be due to the fine-splitting issue for the first year of follow-up. The grouped and individual models without fine-splitting for the first year of follow-up appear to give lower age-standardised estimates. The last row of Table 7.4 for each scenario gives the theoretically true values that would be expected under age-standardisation given that the age-distribution follows a normal distribution. Each of the methods appear to be close to the true value aside from the two rows applied to the yearly split Poisson models.

Figure 7.4 gives a graph matrix that compares the results obtained from each of the methods for each of the first 50 runs of the simulation for Scenario 5. The graphical representation of the results shows that the methods consistently produce different results. The straight line on the plots is a line of equality. If the two compared methods give exactly the same results, the scatter of points would fall exactly on this line. The plot provides a graphical representation of the data contained in Table 7.4 whilst also giving further information on the variation in the estimates from the 500 simulations. Firstly, it is clear the flexible parametric approach and the individual (fine) Poisson model give very similar all-age estimates of relative survival

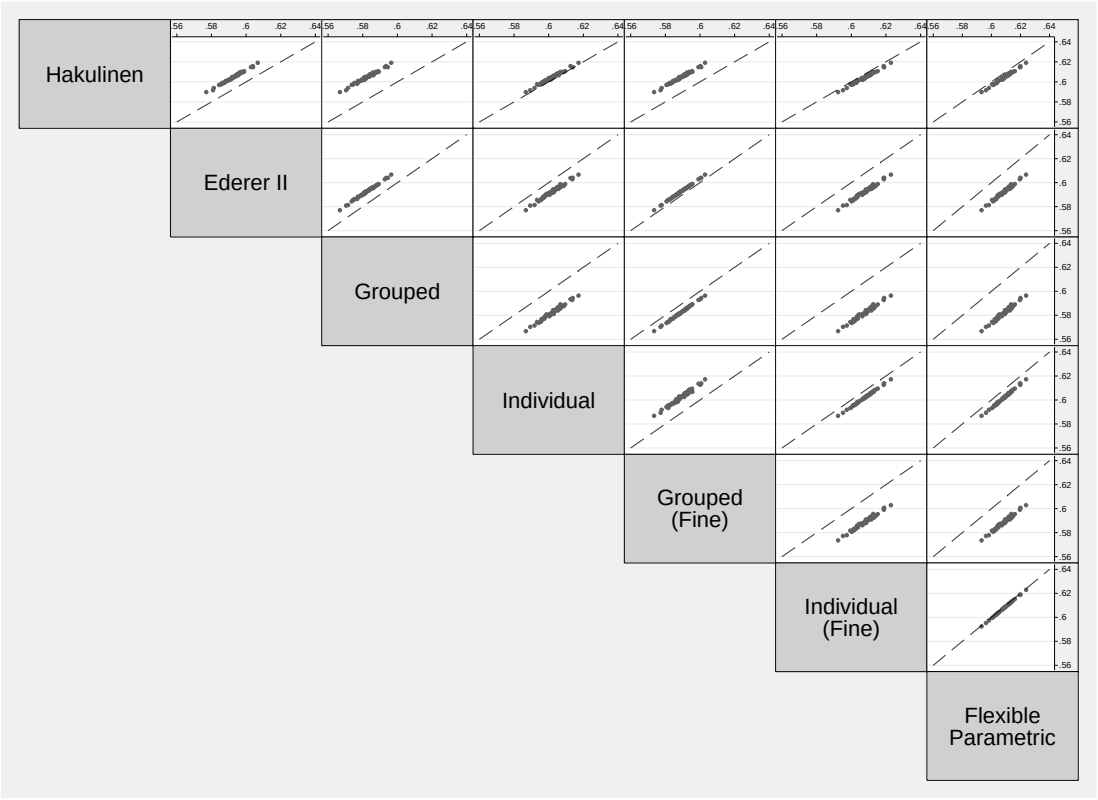

FIGURE 7.4. Scatter Matrix: A pair-wise comparison of each of the all-age estimates from the first 50 simulations carried out for Scenario 5.

for each run of the simulation. Secondly, the grouped Poisson approach provides consistently lower all-age estimates of relative survival than the two individual-level approaches (the flexible parametric approach and the individual Poisson approach). This highlights the fact that the grouped Poisson approach produces biased estimates if the effect of age is ignored. Finally, the last significant feature of the scatter matrix is the difference in estimates that are obtained from finely splitting the first year of follow-up for the Poisson modelling approaches. The finely split approaches for both the grouped and individual level models consistently give higher estimates than the relevant coarsely split estimates.

7.7. Discussion

Under certain circumstances, the different methods for estimating relative survival give vastly different estimates if age is not accounted for. These differences are largely due to the

fact that, for some of the methods, an averaged expected mortality is applied to a group of heterogeneous individuals.

A modelling approach has advantages over the life-table based approaches. Modelling allows greater flexibility for dealing with extra covariates and more complex interactions between covariates. The modelling approaches can also lead to greater precision in situtations where certain assumptions, such as the proportional excess hazards assumption, hold. Another advantage to using a modelling framework with individual data is the ability to include continuous covariates into any analysis.

Within the modelling framework, there is a further choice to make surrounding the use of individual-level or grouped-based data. There are clear advantages to using individual-level data where available, and that the methods proposed in this paper using the individual-level data are theoretically superior than those that do not. It is conceded that in the past there have been computational reasons for not fitting the individual-based models. However, both the individual Poisson model, and particularly the flexible parametric approach now fit sufficiently quickly for even the largest of datasets. The flexible parametric approach has the advantage of not splitting the time-scale into arbitrary windows and, therefore, can fit substantially quicker than an individual Poisson model with finely-split data.

Age-standardisation provides the opportunity to summarise the relative survival experience of a cohort of patients with a single figure. However, if there is large variation between age-groups, information may well be lost by producing only a single summary. Although, the age-standardised estimates can provide age-adjusted estimates for country-wide comparisons, the fact that the difference between countries may vary for different age-groups would not be identified by a single age-standardised figure. This has recently been shown to be the case for an international comparison of breast cancer survival [Møller et al., 2010] where the differences between the Nordic countries and the UK were shown to be confined to early on in follow-up and generally more marked in the oldest age groups.

This chapter has concentrated on comparing these methods using the traditional cohort analysis. However, all the relative survival estimates used here can also be calculated within a period analysis framework [Brenner and Gefeller, 1996]. The issues and differences of the various modelling approaches apply equally to the period analysis setting and are, in fact, just a simple

extension by allowing delayed-entry within any of the models [Smith et al., 2004]. A motivation for the simulations carried out in this chapter was based on the fact that the grouped models that ignore the effect of age are usually applied in the period modelling approach [Brenner and Hakulinen, 2008]. Further work in this area to develop an individual level model suitable for the period-modelling framework could yield more appropriate estimates. This will be covered in the next chapter.

The simulation study was carried out to investigate the main drivers behind the differences observed between the different estimation methods. However, it is conceded that the simulation could have involved more complex scenarios in order to make it more realistic. For example, the simulated datasets that were used had complete follow-up for all patients. This is not usually the case for cancer registry data as censoring is often an inherent part of any analysis. Including the censored data would have made the estimates produced from the Hakulinen and Ederer I approaches different because the Hakulinen approach adjusts for the fact that patients may have heterogeneous follow-up times. However, the reasons uncovered for the differences between the estimates would have been consistent if censoring had been allowed. Another limitation of the simulation was the fact that the age-effects were simulated to apply to the pre-selected categorisation of age. A continuous effect of age would have been more realistic but would have made the modelling and estimation more complicated.

Appropriately summarising relative survival is key for cancer registries and health planning authorities. Individual-based models provide the most appropriate setting for obtaining an effective estimate of relative survival. The methods each provide largely similar estimates providing that they are applied to homogeneous groups of patients but the modelling approaches allow a greater flexibility. In the cases where age is not adjusted for, the individual approaches make more appropriate assumptions than the grouped-based estimates. However, not adjusting for age usually leads to the case of not estimating net survival for the majority of the methods. Of the individual approaches, it is felt that the flexible parametric approach is the most suitable choice due to the fact that an arbitrary splitting of the time-scale is not required. Most importantly, modelling or stratifying by age is vitally important as the lack of homogeneity in the cohort of patients leads to potentially biased estimates of net survival from the relative survival approaches [Rutherford et al., 2011a; Perme et al., 2011].

A recent publication highlights that the net survival can be estimated from the available data [Perme et al., 2011]. Further work is needed to verify this method. However, this appears as though it will be a useful measure as a starting point for the modelling of relative survival in an analogous way to the Kaplan-Meier estimator and modelling overall survival. Unfortunately, the work by Pohar-Perme *et al.* [Perme et al., 2011] was published after the simulation was carried out. Further simulations such as the one carried out in this chapter will determine the effectiveness of this measure in estimating the true net survival.

CHAPTER 8

Modelled Period/Cohort Analysis: Projecting Survival

8.1. Chapter Outline

This chapter further describes period analysis techniques for estimating survival proportions; in particular modelled period analysis. A similar approach in a cohort analysis framework; modelled cohort analysis, is also considered. This modelling framework is used to give up-to-date, and projected estimates of cancer survival. The required method to carry out both approaches is further developed to apply them within a flexible parametric framework. Non-proportional effects of calendar year are considered, and the effect of age is taken into account. An interaction between improvement over time and the age of the patient is also given consideration.

8.2. Introduction

Predicting the burden of cancer in the future requires a projected estimate of survival from cancer. Whereas models for projecting cancer incidence have become popular, there are fewer examples of models for projecting the survival experience of patients. However, there has been extensive literature on providing "up-to-date" estimates of relative survival [Brenner and Hakulinen, 2006a; Mariotto et al., 2006] (introduced in Section 6.7). The survival estimates that are provided by traditional methods can include information on patients that were diagnosed many years before the estimation is carried out [Brenner and Gefeller, 1996]. The main interest for the patient, and health officials, is an up-to-date estimates of the survival proportion, rather than an estimate that applies to patients in the past. That is, the patients want to know what is the survival experience of a patient diagnosed right now.

For most cancer sites, where an improvement over time is generally observed, traditional methods provide an underestimate of the survival probabilities for patients. This also results in the trends and improvements that are obtainable from cancer patient survival data being observed at a significant delay. Period analysis has been proposed as a method that can go

towards correcting these delays by using the survival experience of patients that are closer to the point of interest [Brenner and Gefeller, 1996, 1997; Brenner et al., 2004b] (See Section 6.6). This has become an increasingly popular technique and has been used extensively [Brenner and Hakulinen, 2002; Brenner et al., 2002; Brenner, 2003; Talbäck et al., 2004; Brenner et al., 2011]. On the basis that for the majority of cancers the survival proportion is increasing over time, these estimates would give a closer representation to the "truth" than assuming that the levels of survival remain at a constant when projecting into the future. In order to establish whether or not this is in fact the case, retrospective analyses can be performed. In these analyses, the performance of the various methods in predicting the "future" survival proportions can be compared when, in fact, the survival proportions of the "future" are already a known quantity. This approach has been used to evaluate period analysis methods in previous studies [Brenner et al., 2002; Brenner and Hakulinen, 2008].

Further methods have been suggested to enhance the estimates given by period analysis to give a more accurate, up-to-date estimate of patient survival [Brenner and Hakulinen, 2006a]. These model-based approaches try to allow for the improvements over time to be modelled in order to ensure that the estimate gives a truly up-to-date estimate of long-term survival. A similar approach has been suggested for a more traditional cohort setting for estimating survival [Mariotto et al., 2006].

The approaches that have been suggested are generally used to give up-to-date estimates of survival by making predictions from the last observed time of follow-up. However, it is also possible to use the model-based approaches in order to project further into the future [Verdecchia et al., 2002]. A strong assumption must be made about the continuation of the linear trend in order to make these projections. An example of the projected estimates that can be obtained is given in Section 8.5.

8.3. Up-to-date Survival Estimates

The estimates that are compared are measures of relative, rather than overall, survival. Relative survival is the standard method of analysis for population-based cancer registries because it appropriately accounts for the background risk of death. However, in theory the proposed methods could also be applied to cause-specific or overall (all-cause) survival. The

period analysis approach was introduced in Section 6.6, and modelled period analysis was introduced in Section 6.7.

8.3.1. Modelled period/cohort analysis

Modelled period analysis [Brenner and Hakulinen, 2006a] is becoming a more widely used method for obtaining up-to-date estimates of relative survival, although Brenner himself has been involved in each of the applications of the method to date. The main emphasis for the modelled period technique is to provide an up-to-date estimate of relative survival for patients and health professionals. The basic principles behind period modelling involves fitting a linear trend for attained calendar year within a model in order to attempt to capture the improvements over calendar time and to appropriately account for those in the estimates of up-to-date relative survival [Brenner and Hakulinen, 2006a,b; Brenner et al., 2007; Brenner and Hakulinen, 2009]. A wider diagnosis window than standard period analysis is required to stabilise the estimates of the linear trend [Brenner and Hakulinen, 2006a]. A similar approach has been suggested using a traditional cohort approach (that is, modelling a linear trend for year of diagnosis), rather than using delayed entry techniques [Mariotto et al., 2006]. These two approaches have been compared in a widescale retrospective analysis [Brenner and Hakulinen, 2008] and have been found to give similar results, with modelled period analysis being shown to have smaller standard errors for the estimates. Examples of use of modelled period analysis have recently been appearing in the literature [Gondos et al., 2009a; Brenner et al., 2009b; Pulte et al., 2010].

The modelled period analysis method has been applied using the flexible parametric framework, which was introduced in Chapter 6. The approach is usually applied using a Poisson model, requiring the time-interval to be split to estimate the baseline excess hazard. In essence, the method requires fitting a linear trend for attained calendar year and projecting that to the next observed point in time to get the up-to-date estimate. In order to do this, attained calendar year is simply fitted as a covariate in the flexible parametric model for relative survival (See Section 6.5.4). The approach requires the use of delayed entry techniques, which were described when introducing period analysis in Section 6.6.

That is, the linear predictor, $\mathbf{x}\boldsymbol{\beta}$ for the log excess hazard ratios (from Equation (6.25)), contains a covariate for $year_{att}$.

$$\mathbf{x}\boldsymbol{\beta} = \beta \cdot year_{att}, \tag{8.1}$$

where $\mathbf{x}\boldsymbol{\beta}$ is used as in Equation (6.25)):

$$\ln\{\Lambda(t|\mathbf{x})\} = s\left(\ln(t)|\gamma, \mathbf{k_0}\right) + \mathbf{x}\boldsymbol{\beta}. \tag{8.2}$$

The coefficient β gives a log excess hazard ratio for the effect of a yearly increase in attained calendar year. Attained calendar year varies for each patient over follow-up time, and is therefore a second timescale to consider. To model the second timescale, the data needs to be appropriately set-up. The details of the data set-up are given in the following section. Making a prediction at the final attained calendar year in the available data provides the up-to-date estimate that is required in order to transform to the relative survival scale.

This is in contrast to the Mariotto *et al.* approach [2006] (which has been referred to as modelled cohort analysis), which instead included a covariate for $year_{diag}$:

$$\mathbf{x}\boldsymbol{\beta} = \beta \cdot year_{diag} \tag{8.3}$$

This approach does not require a second timescale to be considered and therefore does not require the data to be further split to perform the analyses. The difference between the two approaches can be seen more clearly from the following section describing the form of the data. Both approaches make projections based upon a linear trend and have been shown to give similar estimates [Brenner and Hakulinen, 2008]. In the same comparison, it was shown that modelled period analysis often provided lower standard errors for the projected estimates than the approach using year of diagnosis.

8.3.1.1. *Form of the data*

To carry out the modelled period analysis approach using flexible parametric modelling, the data must be in the appropriate form. The important distinction is between the two timescales of interest; time since diagnosis, and the attained calendar year value. The dataset must be set up so that there are splits for each new value of attained calendar year, where there is a split for every half-yearly interval across the 5-year window. This results in each patient having multiple rows of data. An example of the format is given in Table 8.1. The splitting for the second timescale is required as the available software only allows a single timescale to be continuous.

Table 8.1 shows the key variables for 3 patients in the dataset. The data relate to a window from 1995-1999. The first patient is diagnosed in 1993, but does not become at risk until the beginning of the window at 1995. Consequently, the start time variable t_0 takes a value of 1.58 for the first row of data for patient 1. That is, the patient's entry to the study was delayed until the start of the window. The attained year for this row is 1995, and the value used in the model for $Year_{att}$ is 0. A value of 0 is used instead of 1995 so that the baseline excess hazard has a direct interpretation. The next row of data for patient 1 highlights that the data has been split into half-yearly intervals. Therefore, the 2$^{\text{nd}}$ row of data for patient 1 relates to the last six months of 1995. However, the patient dies during the last 6 months of 1995 and therefore the survival time is curtailed, and the death indicator variable d takes a value of 1. The $Year_{att}$ variable takes a value of 0.5 for this interval.

Patient 2 in Table 8.1 is diagnosed within the window, in the first half of the year 1995. Therefore, The $Year_{att}$ value used in the model is 0 for the first row of data. Patient 2 dies in the 2$^{\text{nd}}$ half of the year 1996; this means a value of 1 for d and 1.5 for $Year_{att}$ in their final row of data. In this case, the patient began their follow-up at the very start of the window so their survival time t and $Year_{att}$ follow the same pattern through follow-up.

Patient 3 was diagnosed in the latter half of 1997. Therefore, their survival time begins at 0 in 1997. However, the $Year_{att}$ value for their first row of data is 2.5. This is because the attained calendar year is over half way through the window even though the patient has just been diagnosed. This highlights the difference between the two timescales of interest; the survival time from diagnosis, and the attained calendar year for a given patient. Patient 3 is censored at the end of 1999 due to that being the end of the window of interest.

Figure 8.1 shows the follow-up experience for the 3 patients in Table 8.1 graphically. From Figure 8.1, it is clear to see the two different timescales that are being analysed as part of the model. The first time-scale is indicated by the length of the line that falls within the five year window; this ignores where the line starts or finishes. The second timescale takes into account where the patient has reached in terms of calendar time; the attained calendar year, which is split into half-yearly intervals.

8.3.2. Non-proportional effect of calendar year

The main emphasis for period analysis relies heavily on the fact that improvements in survival

Pat. ID	Age. Diag.	Year Diag.	t_0	t	Attained Year	d	$Year_{att}$
1	46	1993	1.58	2.08	1995	0	0
1	46	1993	2.08	2.37	1995.5	1	0.5
2	71	1995	0	0.5	1995	0	0
2	71	1995	0.5	1	1995.5	0	0.5
2	71	1995	1	1.5	1996	0	1
2	71	1995	1.5	1.54	1996.5	1	1.5
3	75	1997	0	0.33	1997.5	0	2.5
3	75	1997	0.33	0.83	1998	0	3
3	75	1997	0.33	1.33	1998.5	0	3.5
3	75	1997	1.33	1.83	1999	0	4
3	75	1997	1.83	2.33	1999.5	0	4.5

TABLE 8.1. Attained Calendar Year

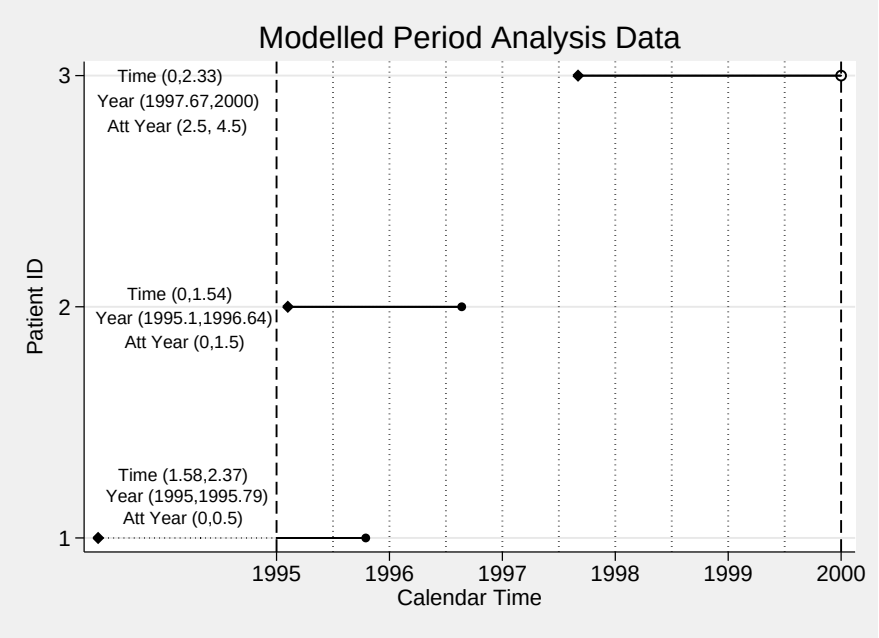

FIGURE 8.1. Figure showing the survival experience for the 3 patients.

are likely to happen in the earlier years of follow-up, whereas modelled period analysis makes the assumption that there is no difference in terms of when the improvements occur with relation to follow-up. In period analysis, the patients whose date of diagnosis falls close, or within, the period window are used to define the follow-up experience for early on in follow-up. That is, period analysis creates a hypothetical new cohort of patients that are as recent as possible in terms of their follow-up time contribution. Modelled period analysis ignores the fact that there may well be a differential effect of calendar year for different follow-up times. The model can

be further developed to test whether this assumption is reasonable by fitting an interaction between attained calendar year and follow-up time. The flexible parametric framework allows spline terms to be fitted in cases when there are non-proportional excess hazards. However, for simplicity it was decided to make a crude split of follow-up time. An effect of calendar year was estimated for those patients in their first year of follow-up, whilst a second effect of calendar year was estimated for longer follow-up times. It is anticipated that the improvements over calendar time will be seen in the first year of follow-up, and lesser improvements over time will be seen for the latter years of follow-up.

That is, the linear predictor for the log excess hazard rate, $\mathbf{x}\boldsymbol{\beta}$, contains a covariate for $year_{att}$ for each of the two follow-up periods, which can be as expressed as:

$$\mathbf{x}\boldsymbol{\beta} = \beta_1 \cdot year_{att}[fu_{<1}] + \beta_2 \cdot year_{att}[fu_{>1}] \tag{8.4}$$

8.3.3. Adjusting for age

As highlighted by the simulation results given in the previous chapter, it is vital to adjust for age when estimating relative survival. In the papers that motivate the use of period analysis, and particularly modelled period analysis, an adjustment for age is not performed, and the "all-age" estimate is given. As highlighted in the previous chapter, relative survival does not give an estimate of net survival when age is not accounted for because the background, and cancer-specific, mortality often depends heavily on age. In the analyses carried out in this chapter, age-standardised estimates will be given in the comparisons. Five standard age-groupings will be used ($\leqslant$44, 45-54, 55-64, 65-74, $\geqslant$75) and internal weights for the cohort of patients' age-distribution will be used to calculate the age-standardised estimates (as detailed in Section 7.3.5).

That is, the linear predictor, $\mathbf{x}\boldsymbol{\beta}$, contains a covariate for $year_{att}$ and a categorical variable for $agegroup$:

$$\mathbf{x}\boldsymbol{\beta} = \beta \cdot year_{att} + \sum_{j=1}^{5} \beta_j \cdot agegroup_j, \tag{8.5}$$

where $agegroup_j$ takes the form: $agegroup_1$ is an indicator variable which takes the value 1 if the patient is in the youngest age-group ($\leqslant$ 44) and 0 otherwise, up until $agegroup_5$ which is an indicator variable for the oldest age-group ($\geqslant$ 75). The model makes the proportional

excess hazards assumption for the effect of age-group; it is possible to relax this assumption, but proportional excesss hazards have been assumed for the analyses in Section 8.4.

From this model, it is possible to obtain estimates of survival for each age-group separately, and then use these to calculate the age-standardised estimate. In Section 8.5 and Chapter 9, age will be modelled continuously using splines rather than using a categorical variable for age. It is necessary to obtain age-specific estimates to obtain appropriate estimates of prevalence, even if an estimate of prevalence for all ages combined is required. Interest also lies in age-specific prevalence of a given disease and age must be appropriately modelled for both survival and incidence if these estimates are desired.

8.3.4. Retrospective analysis

A comprehensive retrospective analysis of the Finnish Cancer Registry data was carried out in order to assess the predictive ability of each of the measures. The Finnish cancer registry data provides one of the longest available datasets for such an analysis, whilst also being renowned for its accuracy and completeness. As well as comparing the modelled period analysis approach to the approach using a non-proportional effect for follow-up time, it was decided that other measures should also be included in the comparison. A traditional cohort analysis approach (referred to as the complete approach by Brenner *et al.*) was included to compare what usually would be calculated. Also, as period analysis has become an increasingly popular tool, two further period analysis estimates were also included in the comparison; one with a 5-year window, and the second with a narrower 2-year window. To summarise, the six compared methods are:

(1) Modelled period analysis (5-year window).

(2) Modelled period analysis with non-proportional effect of calendar year (5-year window).

(3) Traditional cohort approach.

(4) Period analysis approach (5-year window).

(5) Period analysis approach (2-year window).

To use the long time-series of data in the comparison, 5 year relative survival estimates were obtained for a range of calendar years. The following analysis explained for the start year of 1970 was repeated in yearly increments up until the year 2000. To start, an age-standardised

estimate of relative survival was obtained from a flexible parametric model, using both a cohort and period approach, for patients diagnosed between 1970 and 1974. Then a narrower period window estimate for the last 2 years of this 5 year diagnosis window was calculated. A flexible parametric model including a covariate for attained calendar year over the diagnosis window was also estimated. Predictions from this model were made for the last attained calendar year (1974). Finally, the model detailed with a non-proportional effect for calendar year over follow-up was used (with the single split after a single year of follow-up); again, predictions were made for the last attained year.

The analysis that is explained above starting in the year 1970 was repeated in yearly increments up until the year 2000. When making the comparisons, an average over the 31 estimates for the range of time-points was calculated.

8.3.5. What to use as the comparitor?

A decision needed to be made about what should be compared to the projected estimates in order to validate the estimates of survival. To assess the up-to-dateness of the obtained estimates, a standard relative survival estimate was also obtained for an interval that was centred on the end point of the interval used to fit the models. That is, if the model was fitted to patients diagnosed between 1990 and 1994, then the truth was calculated for a 5-year window centred on the end of 1994; patients diagnosed in the last 6 months of 1992, up until patients diagnosed in the first 6 months of 1997. This will be used as the observed truth, and will then be compared to the various estimates using three different measures; the mean difference, the absolute mean difference, and the mean-square difference. These three measures will highlight the different features of the methods to estimate relative survival. Methods that perform well for all three measures are to be considered a good estimator of up-to-date relative survival.

8.3.6. Data

The datasets that were used to perform the analyses were provided by the Finnish Cancer Registry. Ten cancer sites were chosen, and the full extent of the length of the data was exploited. Common cancer sites were selected to provide a sufficiently large dataset to peform the modelling, and to allow the appropriate power to detect the interaction with follow-up time. The population of Finland has grown from just over 4 million in the 1950s, to around 5.3 million by the end of 2008. This highlights that the number of cancer cases observed over this period

would not be on the scale of other countries, and further justifies the selection of only the most common cancer types. The ten chosen sites are bladder, breast, colon, lung, non-Hodgkin's lymphoma, pancreas, prostate, rectum, stomach and thyroid gland.

Lung cancer was selected as an example to highlight the issue of ignoring the potential interaction with follow-up time. The standard period modelling techniques that have been applied to date ignore this interaction and fit a weighted average of the two gradients; with the weights being based around the time of deaths. However, the true interest lies in whether or not the model still manages to provide a reasonable estimate for the up-to-date prediction.

8.4. Results

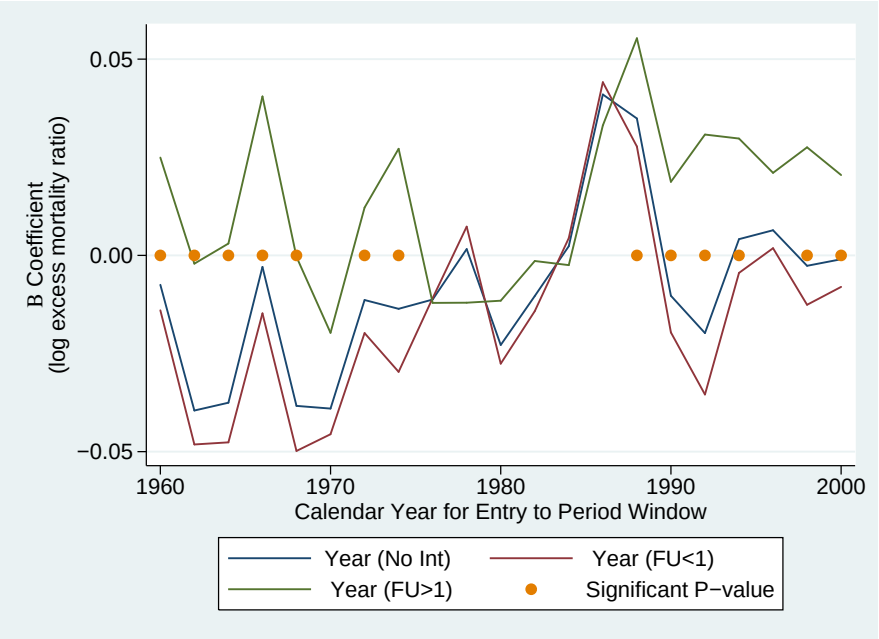

FIGURE 8.2. Values of the three different log-excess hazard ratios over time for lung cancer.

Figure 8.2 compares the coefficient values from the model with a proportional effect of calendar year compared to the model that uses a split for follow-up time after a year. The values given in the figure are log excess hazard ratios for a unit increase in calendar year. Therefore, a value that is less than 0 suggests that there has been an improvement over calendar time, whereas a positive value suggests that the mortality is increasing over calendar time for the interval under study. The figure relates to lung cancer in Finland, and highlights that over time,

there is usually a stronger improvement for patients in the first year of follow-up compared to later years. The scatter plot (at the value of 0 on the figure) highlights when the inclusion of the interaction with follow-up time provides a significantly better fitting model according to the likelihood ratio test at the 5% level. Standard modelled period analysis assumes that this interaction with follow-up time does not exist.

It is clear that during the 1960s and 1970s there are improvements in terms of relative survival from lung cancer. This can largely be attributed to introductions of new treatment strategies involving chemotherapy. However, during the 1980s and 1990s, there appears to be less improvement. In the 1990s, there is a significant time-dependence for the effect of calendar year and follow-up time. This appears to indicate a period of improvement in terms of patients in the first-year of follow-up, but of a worsening over time for those with a longer-term follow-up. A Europe-wide review of the epidemiology of lung cancer found similar trends over calendar time across Europe; despite the inherent differences between the countries [Janssen-Heijnen and Coebergh, 2003].

As has been recently observed in an international comparison, improvements in relative survival for lung cancer are often largely made early in follow-up [Coleman et al., 2011]. This is verified by the results in Figure 8.2. There is some random noise in the value for the linear effect of calendar year over time; however, it is clear that in times when there are improvements, the improvements can be more largely attributed to early on in follow-up. In fact, lengthening the life of patients early in follow-up can actually have a detrimental effect for the coefficient for the later years of follow-up. This occurs when the life is lengthened through improvements in treatment, but the cancer is not fully cured.

Figure 8.3 shows a selection of the compared estimates for breast cancer across the entire analysis period. The plotted estimates are the 5-year traditional age-standardised estimates of relative survival by the various estimation approaches. It is clear that the standard cohort approach gives lower estimates, as well as showing changes in the trend at a delay. This is exactly what is expected considering the data that is used in order to obtain this estimate. The modelled period approach gives a closer estimate to the truth than the other measures. Using a linear trend for attained calendar year, and predicting at the final year, appears to give a closer indication of what is observed when comparing to the true estimates from the later

window. The two-year window for the period approach appears to give inflated estimates of age-standardised relative survival in the case of breast cancer over the vast majority of the time range.

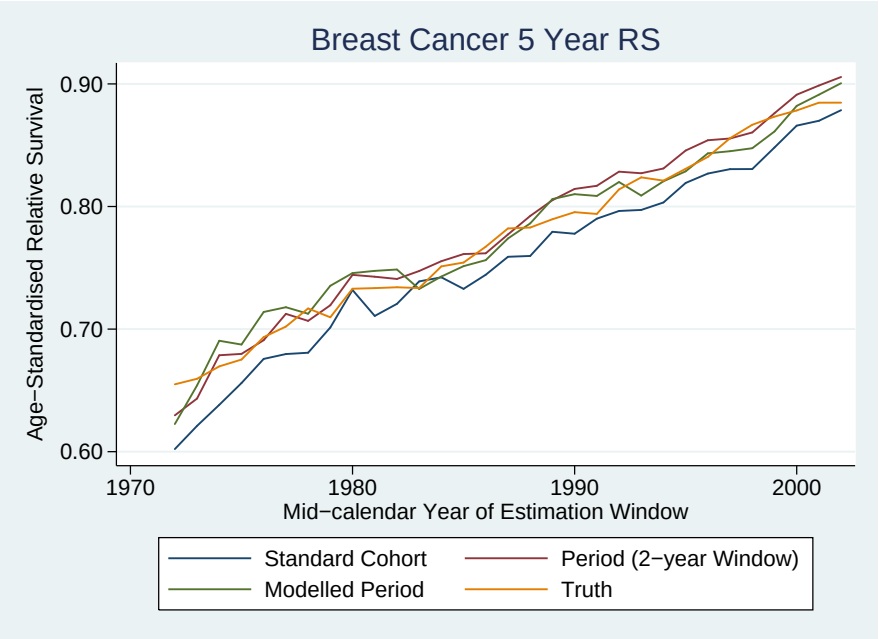

FIGURE 8.3. Age-standardised 5-year relative survival estimates over time for breast cancer.

Table 8.2 shows the results of the comparison of the various methods for obtaining an estimate of 5-year age-standardised relative survival for the 10 cancer sites. In this case, the results are compared by summarising the mean of the difference between the observed true estimate and the various approaches for each time period over the 31 year period. A positive value indicates that on average the compared approach underestimates the "truth" whereas a negative value indicates that an overestimate is obtained on average over the 31 estimates. As seen in the breast cancer example portrayed in Figure 8.3, the complete approach (standard cohort) tends to underestimate the later observed "true" value. For lung cancer, the complete approach appears to perform particularly well. This is because there have been very few improvements over time in terms of the relative survival estimate for lung cancer.

The period approach (using the full 5-year window) performs a little better in terms of the mean difference estimates for the 10 cancer sites, but only fractionally. Using such a wide window for the period approach negates the effectiveness of the approach. However, using

Cancer	Complete	Period	Period (two)	Period Mod	Period Mod (sp)
Bladder	0.860	1.056	-2.433	-1.158	-1.176
Breast	1.924	2.216	-0.610	-0.284	-0.099
Colon	1.933	0.626	-1.562	-1.820	-1.584
Lung	-0.007	-1.130	-1.378	-1.322	-0.884
Non-Hodg	2.362	1.790	-0.687	-0.832	-1.274
Pancreas	0.127	-0.308	-0.266	-0.529	-0.224
Prostate	3.036	3.938	1.283	0.353	1.168
Rectum	1.641	0.760	-1.665	-1.163	-1.242
Stomach	1.146	0.050	-1.063	-1.221	-0.802
Thyroid	1.742	0.014	-3.645	-1.387	-1.471

TABLE 8.2. Average value for the calculated difference across the 31 time intervals. Difference calculated between the observed truth and the estimate from the method (Age-standardised 5 year RS).

Cancer	Complete	Period	Period (two)	Period Mod	Period Mod (sp)
Bladder	2.167	2.028	2.824	2.538	2.737
Breast	1.959	2.216	1.067	1.140	1.158
Colon	2.012	1.063	1.950	2.280	2.133
Lung	0.540	1.130	1.378	1.339	0.915
Non-Hodg	2.630	2.101	2.005	2.500	2.663
Pancreas	0.304	0.434	0.405	0.624	0.362
Prostate	3.544	4.291	2.599	2.626	2.777
Rectum	2.361	2.087	2.014	2.206	2.385
Stomach	1.528	1.208	1.618	1.405	1.276
Thyroid	2.368	1.689	3.833	2.544	2.558

TABLE 8.3. Average value for the absolute difference across the 31 time intervals. Difference calculated between the observed truth and the estimate from the method (Age-standardised 5 year RS).

the two-year window appears to suggest that on average an overestimate is produced for the majority of the 10 cancer sites. This could be due to the truth that was chosen as the comparitor, which contained data from both the past, and near future.

The results of the modelled period analysis approach are contained in the 5^{th} column of Table 8.2. The modelled period analysis approach appears to perform better on average (in terms of the mean difference) than the other compared methods. Using the categorical split for follow-up time for the period modelled approach, which is given in the next column, does not appear to improve the estimation despite often being a better model for some cancer sites. Although, the split for follow-up time is certainly evident for some of the cancer sites, it appears that assuming an average yearly trend over the entire range of follow-up produces equally good estimates.

Cancer	Complete	Period	Period (two)	Period Mod	Period Mod (sp)
Bladder	6.476	5.918	11.323	9.993	11.277
Breast	5.077	6.144	1.510	1.861	1.905
Colon	5.973	2.242	5.456	8.113	6.672
Lung	0.420	1.682	2.252	2.290	1.243
Non-Hodg	9.769	7.076	5.823	8.503	10.110
Pancreas	0.132	0.258	0.235	0.503	0.179
Prostate	16.853	27.705	9.714	9.492	10.436
Rectum	7.351	5.622	6.364	6.577	7.840
Stomach	3.947	2.322	3.857	3.209	3.012
Thyroid	11.002	6.068	23.659	14.112	15.941

TABLE 8.4. Average value for the mean-squared difference across the 31 time intervals. Difference calculated between the observed truth and the estimate from the method (Age-standardised 5 year RS).

Table 8.3 shows the equivalent results using a measure of absolute difference between the methods and the later "truth". Whereas in Table 8.2 an overestimate could be cancelled out with an underestimate in the preceding year, that is not the case when using an absolute measure. However, the results contained in Table 8.3 show a similar pattern to those in Table 8.2. As has been seen in similar evaluations [Brenner and Hakulinen, 2006a; Gondos et al., 2009a], the modelled period approach appears to perform well across a number of cancer sites. The approach proposed using a time-dependent effect of calendar year ("Period Mod (sp)") does not appear to perform better than the standard modelled period approach. However, there does appear to be an improvement for lung cancer, which verifies the findings in Figure 8.2.

Table 8.4 shows the final comparison that was made between the later observed truth and the compared estimates. Using the mean-squared difference between the estimates and the truth will punish the methods that produce more varied estimates around the observed true value. Pancreatic cancer and lung cancer have very low 5-year relative survival that has shown little improvement, and it appears that this results in the methods producing fairly stable estimates. The modelled period analysis approach also performs well using this metric. It is clear that fitting survival models that contain a linear effect of calendar year can be used to produce good estimates of short-term projections for survival.

8.5. Projecting Survival

In the previous section, up-to-date estimates of 5-year age-standardised survival were compared. These estimates are predictions at the end of the available data rather than providing estimates of future survival. However, the outlined methodology is equally capable of providing projected estimates of survival. A strong assumption surrounding the continuation of the linear trend for attained year (or year of diagnosis) is required to make the projections.

The necessary model for the projections are equivalent to the models set out in Section 8.3.1. In the previous section, the modelled period analysis approach was used for the comparison. The Mariotto *et al.* approach [2006] (modelled cohort analysis) would have been equally valid, and the two approaches have been shown to give similar estimates over a wide range of cancer sites and an extensive time period using the Finnish registry data [Brenner and Hakulinen, 2008]. The modelled cohort analysis approach will be used in the following sections across a number of examples to show the results when projecting relative survival. However, a comparison of the estimates obtained from both the modelled cohort and modelled period approaches will also be given.

8.5.1. Extensions to the Models

The models in this section are given from the perspective of modelled cohort analysis. They could equally be applied within the modelled period analysis framework by exchanging $year_{att}$ for $year_{diag}$ and ensuring the data was in the appropriate form. To develop the models contained in Section 8.3.1 further, it is possible to include spline terms in the model rather than categorical covariates. That is, the model contains a covariate for $year_{diag}$ and a spline function for the effect of age:

$$\mathbf{x}\boldsymbol{\beta} = \beta \cdot year_{diag} + S_{df_a}(age), \tag{8.6}$$

where the subscript df_a is used to denote the degrees of freedom for the spline function, and age refers to the age at diagnosis.

This model assumes that the effect of calendar year is the same across the entire age range. It is possible to relax this assumption by fitting an interaction between the two terms:

$$\mathbf{x}\boldsymbol{\beta} = \beta \cdot year_{diag} + S_{df_a}(age) + \beta_1 year_{diag} \cdot S_{df_a}(age). \tag{8.7}$$

If a large value is selected for the degrees of freedom for the spline term for age, this model may result in the interaction term being over-fitted. A similar problem was discussed in Chapter 2, where a "reduced" set of splines was used for the interaction between sex and age. Similarly, using a reduced set of splines for time-dependent effects in flexible parametric survival models was discussed in Chapter 6. Following a similar argument, a reduced set of splines can be used for the interaction term whilst using the more complex function for the effect of age:

$$\mathbf{x}\boldsymbol{\beta} = \beta \cdot year_{diag} + S_{df_a}(age) + \beta_1 year_{diag} \cdot S'_{df'_a}(age), \tag{8.8}$$

where $S'_{df'_a}(age)$ is a spline function for age at diagnosis with fewer parameters; with degrees of freedom $df'_a < df_a$.

Finally, it is also possible to assume that the effect of calendar year of diagnosis is not linear by using a second spline function:

$$\mathbf{x}\boldsymbol{\beta} = S_{df_y}(year_{diag}) + S_{df_a}(age). \tag{8.9}$$

Allowing the effect of calendar year to be non-linear could possibly lead to erratic projections when projecting into the future. A similar problem was encountered in Chapter 4 when projecting incidence, and was counteracted by moving the boundary knot within the range of the data to enforce a linear function. Therefore, this model may well perform poorly when making projections.

8.5.2. Example

Firstly, a comparison of the two modelling approaches is considered for the example datasets. Figure 8.4 shows the excess hazard ratio effect of age from the two modelling approaches (modelled period analysis and modelled cohort analysis) for the Finnish colon cancer data. The age term is fitted with a spline function with 10 degrees of freedom, and 5 degrees of freedom are used for the underlying baseline excess hazard. The age functions are given relative to age 60 by subtracting the relevant value of the spline function at age 60 from each of the spline terms. The two lines overlay exactly and it is not possible to distinguish them. Also reported on the graph is the excess hazard ratio for the effect of attained calendar year (0.9696), and year of diagnosis (0.9708). It can be seen that the estimates are very similar and therefore the two models should give almost equivalent projections. Figure 8.5 shows the prediction for a 60

year old in 1995 from the two approaches. It can be seen that similar estimates are given from either modelling approach.

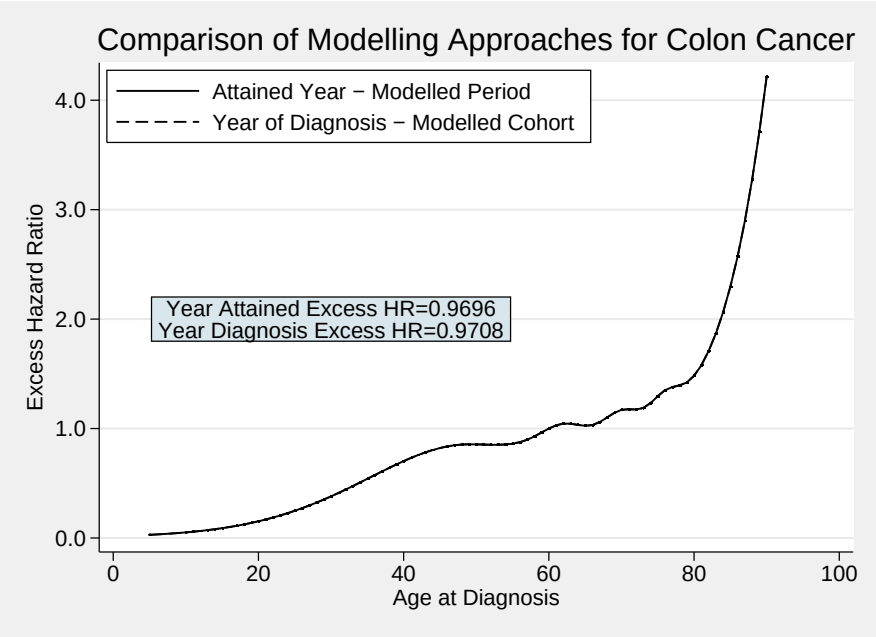

FIGURE 8.4. Comparison of the estimates from the two modelling approaches. Colon cancer data from 1985-1995, excess hazard ratio for the effect of age.

Figures 8.6 and 8.7 show the equivalent predictions for a 60 year old in 1995 for both the lung cancer dataset, and the breast cancer dataset respectively. Again, it is clear that similar estimates of relative survival are obtained from the two approaches. On the basis that the setting up of the data is simpler for the modelled cohort approach and similar estimates are obtained from both approaches, the modelled cohort approach is used for the remainder of the example for exploring the extensions described in the previous section.

The extensions to the models for modelled cohort analysis were firstly applied to Finnish colon cancer data. Figure 8.8 shows the excess hazard ratio for age for colon cancer patients diagnosed between 1985 and 1995. The two lines relate to the models described in Equations (8.5) and (8.6), one with a categorical effect of age and the second with a spline function with 10 degrees of freedom to describe the effect of age at diagnosis. It is clear that the spline model provides a smooth representation of the categorical effect. The factor model for age gives an average effect over wide age categorisations. The spline models smooths over finely split data in

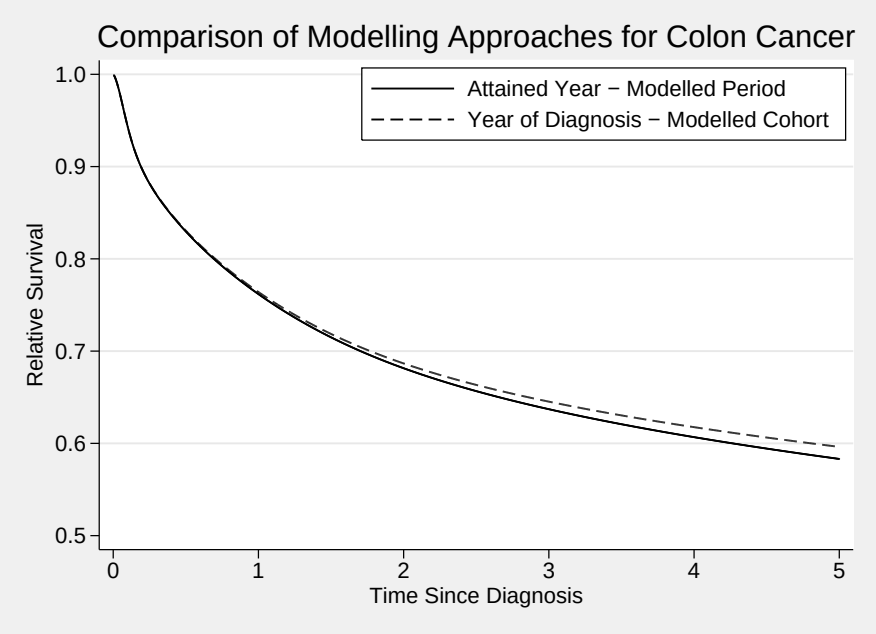

FIGURE 8.5. Comparison of the estimates from the two modelling approaches. Colon cancer data from 1985-1995, prediction at 1995 for a 60 year old patient.

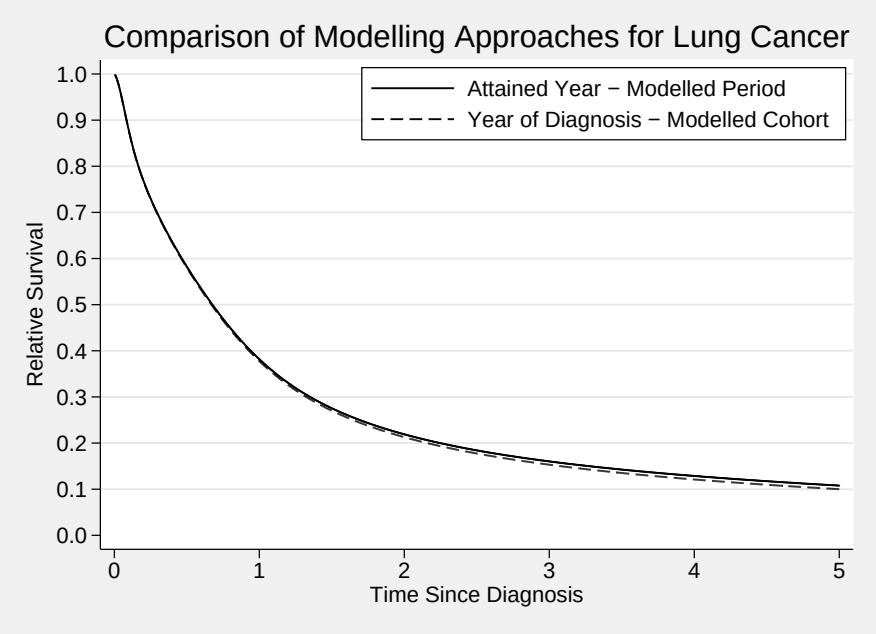

FIGURE 8.6. Comparison of the estimates from the two modelling approaches. Lung cancer data from 1985-1995, prediction at 1995 for a 60 year old patient.

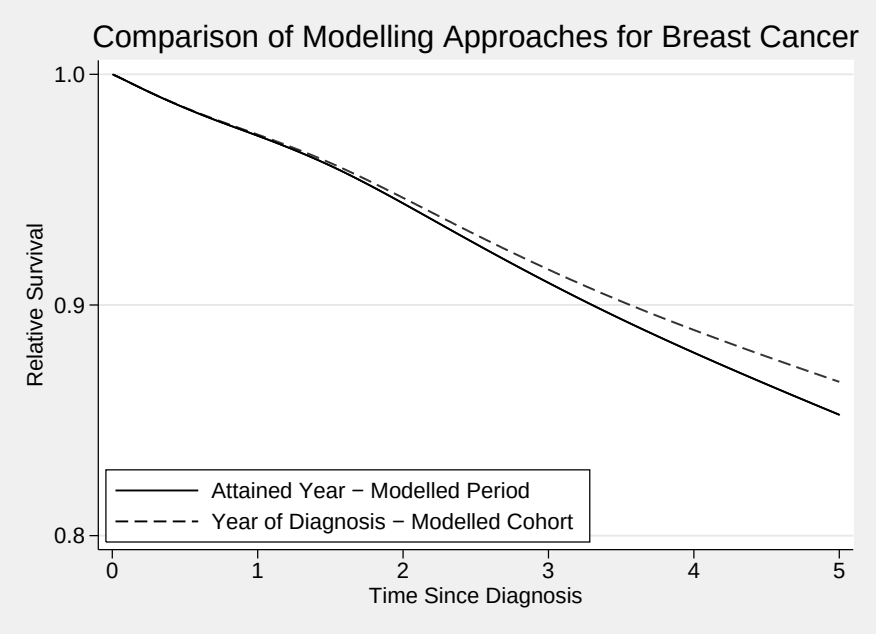

FIGURE 8.7. Comparison of the estimates from the two modelling approaches. Breast cancer data from 1985-1995, prediction at 1995 for a 60 year old patient.

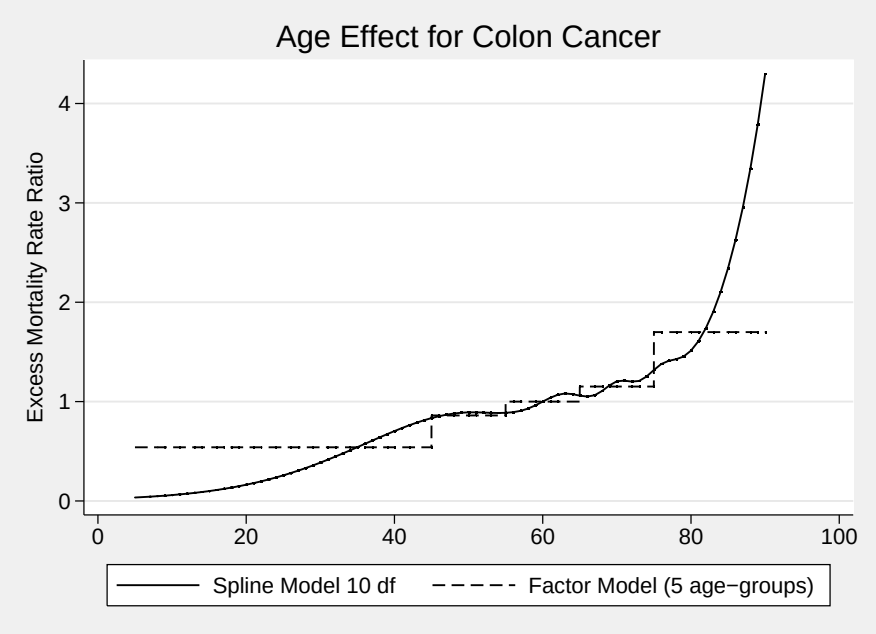

FIGURE 8.8. Age Effect for Colon Cancer. Comparison of models using splines to the factor model for the effect of age.

order to provide a continuous representation of the effect of age. This is similar to the approach for incidence models advocated in Chapter 2.

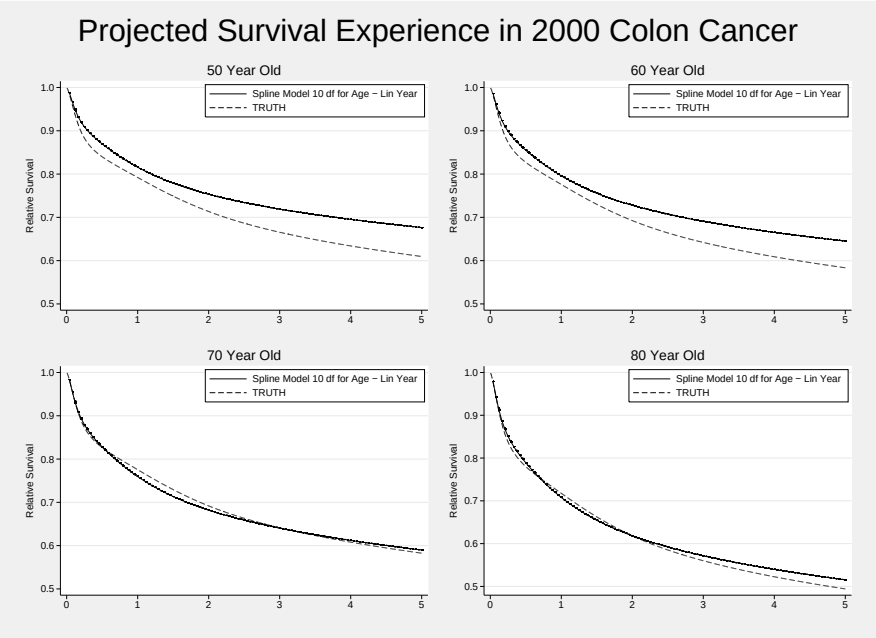

FIGURE 8.9. Projections for Colon Cancer Patients in 2000 for 4 specific ages, based on data 1985-1995.

Figure 8.9 shows the projected survival experience in the year 2000 using colon cancer data for years of diagnosis between 1985 and 1995 (that is, a 5-year projection into the future). The projections are compared to the true estimates of relative survival for patients diagnosed in the year 2000. The model fitted to the data is described in Equation (8.6), with 10 degrees of freedom used for the spline function for age at diagnosis. The projections are shown for 4 representative ages; the projections are different for each value of age at diagnosis because of the fact that a spline function is used for age. Overestimates are produced for the younger two ages (50 and 60), but the projected estimates perform well for the 70 and 80 year old patients. The projections are based upon the assumption that the linear trend over diagnosis year observed in the 1985-1995 diagnosis window is continued until the year 2000. A second assumption is that the effect of age observed during that window also holds for the projection year. On the basis that only the main effect of age and year of diagosis have been used in this model, a further assumption is that the improvement over calendar time is the same across all

ages. This assumption appears to be reasonable for colon cancer data as the inclusion of an interaction term did not produce a statistically significant improvement at the 5% level.

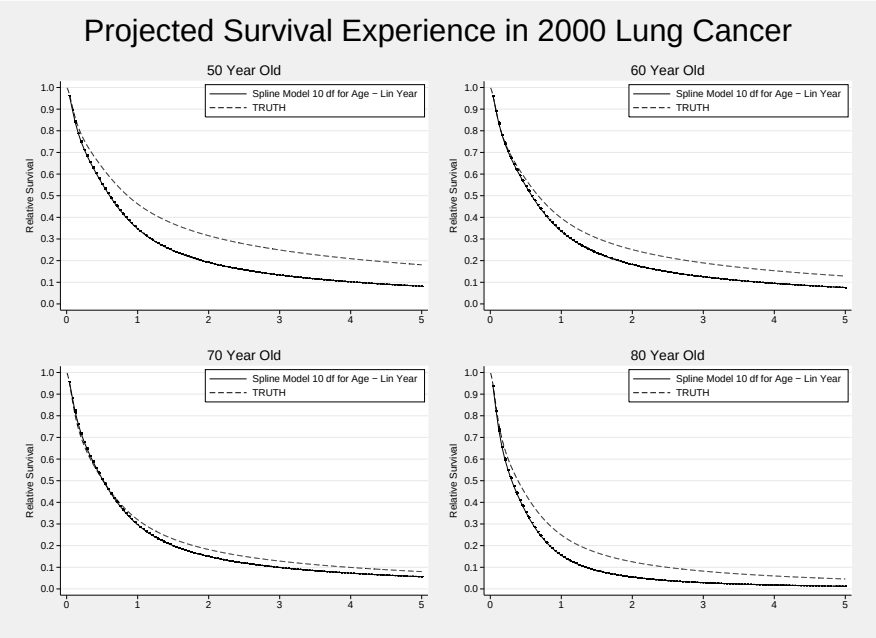

FIGURE 8.10. Projections for Lung Cancer Patients in 2000 for 4 specific ages, based on data 1985-1995.

Figure 8.10 shows the results for lung cancer over the same diagnosis period, for the same projection year of 2000. In this case, underestimates are produced for each of the 4 selected ages. This result can be explained by referring back to the information contained in Figure 8.2. For the diagnosis period of interest (1985-1995), there is evidence that the mortality is increasing over calendar year for that period. For the year 2000, it appears that the effect of calendar year has reduced in comparison to the earlier years. This difference is expressed in the results contained in Figure 8.10. Using a 10 year-window of diagnosis can result in a more stable estimate for the effect of year of diagnosis. However, the data used for the projection is consequently less recent, and this can result in poorer projections of relative survival.

Figure 8.11 again shows the projections for lung cancer for the year 2000 but instead using a shorter diagnosis window (1990-1995). This window includes years of diagnosis that are similar to that of the year 2000 (see Figure 8.2). This leads to the projected estimates of relative survival giving more similar estimates to those observed in the year 2000. As with

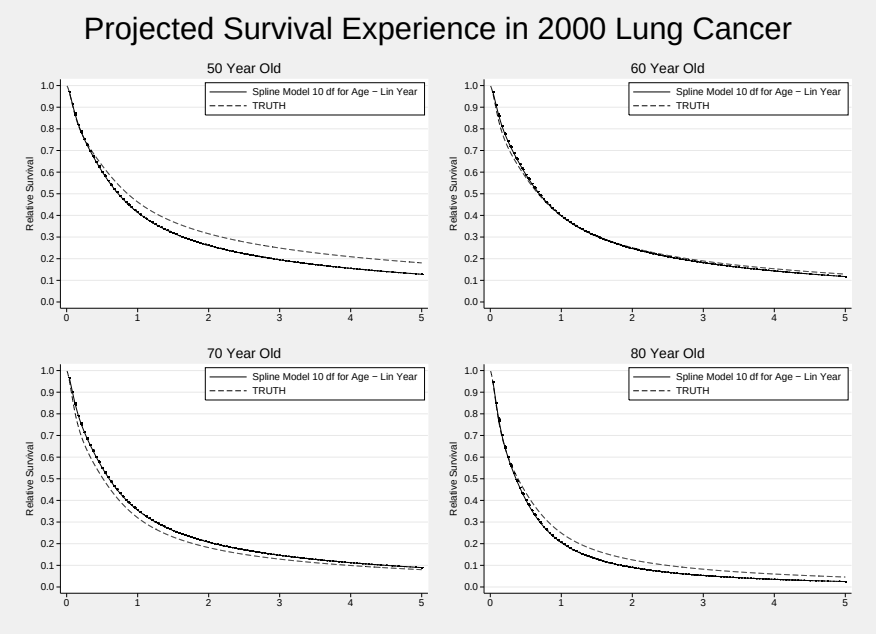

FIGURE 8.11. Projections for Lung Cancer Patients in 2000 for 4 specific ages, based on data 1990-1995.

the projection made in Chapter 4 for incidence, there is a balance between "recentness" and stability that needs to be met in order to obtain relevant projected estimates.

Figure 8.12 shows longer-term projections for breast cancer in 2005 based on a model using patients diagnosed between 1985 and 1995. A further comparison is added to the figure from a model using an interaction between year and age of diagnosis. The estimates from the interaction model given in Figure 8.12 are obtained from the model described in Equation (8.8), with 10 degrees of freedom used for the main effect of age, and 3 degrees of freedom used for the age effect in the interaction term.

It is clear that the model with the interaction term included performs better; particularly for the older patients. The interaction allows the effect of year of diagnosis to vary across the age-range. The improvements made for breast cancer patients appear to be consigned to the younger patients, and assuming the same level of improvement for an 80 year old leads to a severe overestimate of relative survival. The interaction between age and the improvement over calendar time could be a logical conseqence of different treatment strategies for patients of different ages.

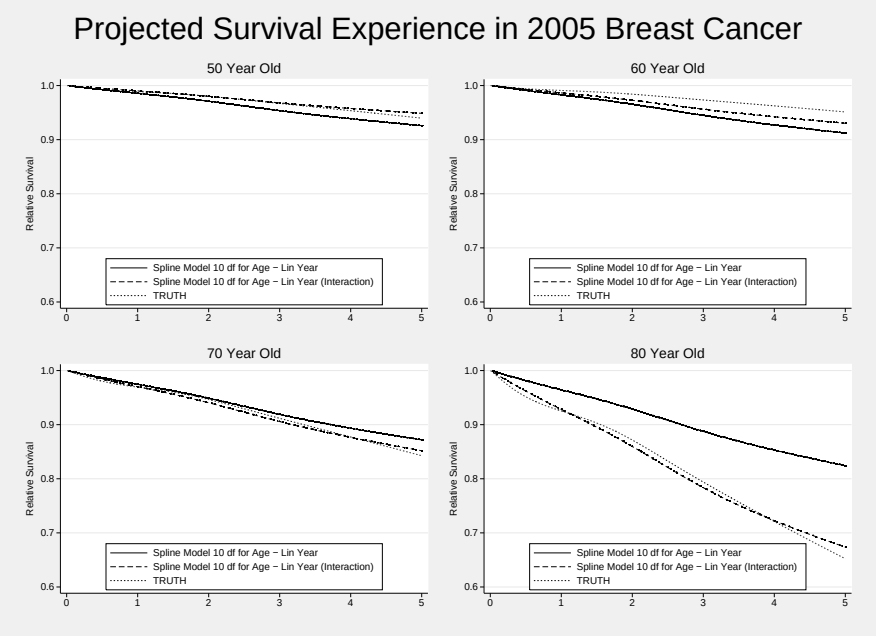

FIGURE 8.12. Projections for Breast Cancer Patients in 2005 for 4 specific ages, based on data 1985-1995.

8.6. Discussion

In this chapter, the modelled period and modelled cohort approaches has been applied within the flexible parametric modelling setting. Poisson models have previously been used to apply the modelled period analysis approach. Using splines to describe the underlying baseline excess hazard gives an improvement over having to make an arbitrary choice over how to split the timescale. However, in order to deal with the second timescale of attained calendar year for modelled period analysis, splitting was still required for the flexible parametric approach. A second development of the current approach was to introduce a non-proportional effect of calendar year within the modelled period analysis setting. In the final section of the chapter, making projections using the modelled cohort analysis approach was developed. Improvements to this modelling approach were suggested by using an interaction between the effect of age and the linear trend improvement over calendar time.

The results in Section 8.4 show that modelled period analysis appears to perform well as a method for obtaining up-to-date estimates of relative survival. The traditional approaches to estimating relative survival produce estimates that are out-of-date for current cancer patients

and also fail to detect changes in survival trends quickly. Period analysis approaches, and particularly modelled period analysis, appears to correct for these deficiences to some extent.

In order to assess the relative performance of the methods, it was necessary to define a "truth" to make a comparison. The decision over what to use as the "truth" in this retrospective analysis could have a significant effect on the interpretation. It was decided that an appropriate truth would be a standard relative survival estimate from a five-year window that was centred at the point of interest. This method was selected in the hope that it would give a reliable and stable estimate of the relative survival that was relevant to the point of interest by using data that was both "in the future", and the "recent past".

The models that were used to calculate the relative survival are parametric, and use restricted cubic splines to model the shape of the baseline hazard. Models that involve splines are often criticised on the basis that the knot selection is somewhat arbitrary. There has been work to show that the knot selection that is used for these models does not greatly effect the estimates that are produced. The same number of knots were used for the baseline excess hazard for every model in the analysis (4 internal knots) and the knots were defined to be equally spaced on the log of follow-up time for event times. The flexibility that is introduced from fitting models rather than using the life table estimates outweighs the cost of having to decide upon the number of knots to use for these parametric models.

The decision to create a simple interaction with follow-up based on a single split after one year reflects the fact that period estimation places the most emphasis on the most recent year of study in order to obtain an estimate of survival. Other split points were attempted but this did not lead to a significantly better fitting model for the majority of cancer sites. Another avenue that was explored was to fit a collection of spline terms for the effect of follow-up. This again, failed to yield significantly better models but could arguably be a potential extension to the crude splitting of the timescale that has been trialled in this chapter. The message from period analysis in general is that the major improvements for most cancer sites is seen in the early years of follow-up, particularly in the first year of follow-up; the models that have been fitted allow this standard assumption of period analysis to be extended to modelled period analysis.

The interaction with follow-up time often provided a better fitting model to the data, but did not result in better estimates of up-to-date survival. This highlights that even if a model

fits the current data well, it may not necessarily imply that it can be then used to make sensible projections. It is often better to make a simple fit to the data in order to make projections. In order to extrapolate from a model, the model needs to fit the data well, but not so well that it cannot be applied outside the range of the current data. However, if interest lies in only understanding the trends and changes within the range of the data, then the interaction model provides further insight than the simple linear approach.

Using a longer window for the modelled period analysis may well lead to more stable projections. The estimates over a five-year window could potentially be quite "noisy" for some sites, particularly considering the population size in Finland. Fitting a linear trend through five point estimates may well lead to an over-, or under-estimate in some cases. A longer window would include less recent data, but may well lead to a more stable fit for the linear trend. The balance between recentness and stability for making projections is similar to that needed for the incidence projections described in the earlier chapters.

In Section 8.5 the modelled cohort analysis approach was used to give projections of survival estimates rather than up-to-date estimates of survival. Making projections from the modelled cohort and period analysis approaches can be achieved by making out-of-sample predictions for the models. However, as projections are extended to further points in the future, the likelihood that the trend observed in the data window still holds lessens. This is a similar issue that is relevant for any predictions based on data and was also discussed in Chapter 4. Instead of using a categorical effect of age, splines were used in order to be able to give a more continuous estimation of the effect of age. A further extension to the models used in Section 8.4 was the inclusion of an interaction term between the linear effect of year of diagnosis and the spline terms for age at diagnosis. This interaction proved important for the example concerning breast cancer and this is certainly an extension that will need consideration when using the survival projections to make projections of prevalence.

The modelled cohort and modelled period analysis approaches were compared for three cancer sites. The approaches are similar in theory and provide similar projected estimates in the majority of cases. The difference between the two approaches lies in the distinction between fitting a linear trend for a patient's attained calendar year and a patient's year of diagnosis. These often result in similar estimates of projected relative survival. The modelled cohort

approach does not require the splitting of the data to account for the second timescale. Year of diagnosis is a covariate that is fixed for each patient and therefore can simply be included in a standard model without the need to split the data. Therefore, this approach was favoured for the remainder of the exploration of the potential for projecting survival.

CHAPTER 9

Estimating and Projecting Prevalence: Combining Survival and Incidence

9.1. Chapter Outline

In this chapter, the methods developed in the chapters on survival and incidence are combined in order to provide an estimate of prevalence. Using the projection techniques outlined in Chapters 4 and 8, it is possible to provide projected estimates of cancer prevalence as a proxy for the future cancer burden. The improved projection techniques for incidence and survival that have been described, should lead to improved model-based estimates of prevalence. Finnish cancer registry data is used to illustrate the approach.

9.2. Introduction

Prevalence provides a more complete estimate of cancer burden as it combines the information obtained from survival and incidence data into a single measure. Prevalence is commonly defined as the number of people with a disease of interest at a given point in time. The estimates of future cancer incidence are of interest on their own, and provide useful measures of definining the burden of cancer [Bray and Møller, 2006]. The incidence can be used to define the number of new cases, in a specified time period, and therefore projections of this quantity provide information to health planning authorities on the number of patients that will need immediate cancer care in the future. However, a large proportion of cancer patients live well beyond their point of cancer diagnosis and can require treatment over an extended period of time; this will incur a cost which will be a financial burden for the health authorities. Attempts have been made to assess this financial burden in the Nordic countries using prevalence information [Kalseth et al., 2011]. The length of the time for each patient is defined by their follow-up time to death, and the proportion of patients that survive until time-point T is defined by the estimate of the survival proportion. Those cases that are still alive at time-point T have had

a diagnosis of cancer and are considered a prevalent case. Therefore, improvements in cancer patient survival lead directly to an increase in the prevalence of cancer.

9.3. Types of prevalence

Prevalence is used as a general term and in order to be clearer, the following definitions will be used for the analyses carried out and for the discussions that follow.

9.3.1. Total prevalence

Total prevalence is the standard definition of prevalence for a cancer diagnosis. That is, a prevalent case is defined as anyone who is still alive and has had a previous diagnosis of cancer. In this case we assume that the diagnosis of cancer is irrevesible; this has also been referred to as diagnosis prevalence [Capocaccia and De Angelis, 1997]. It is clear how this definition is a direct relationship between incidence (new cases) and the survival (whether a patient is still alive). Total prevalence can be reported as a proportion of the total population, or can be given as the total number of cases in the population satisfying the above criteria. The common criticism of the total prevalence as a measure of cancer burden is that patients who are long-term survivors of their cancer diagnosis (and may well no longer be a burden due to their cancer), are still included in the prevalence estimate. This leads to an inflated estimate of the number of burdensome cases of cancer in the population. The difficulty in providing any other measure than this is distinguishing between cases that are a burden, and those that are not. Cancer registries do not follow-up cancer patients closely enough to make that judgement; the only follow-up information that is usually available is via linkage to the death registry to ascertain whether or not the patient has died.

Figure 9.1 shows a Lexis diagram for a small subset of patients. If an estimate of prevalence is required at the year 2000, the intersection of the follow-up experience of the patients with a vertical line at the year 2000 (as shown) can be used to calculate the total number of prevalent cases of the disease of interest. As described in Section 2.4, the circles indicate the point of diagnosis for a particular patient (incidence), and the diagonal lines are indicative of the time the patient is followed-up (as the patient ages, and calendar time continues); that is, the patient's survival experience. If a patient is still alive at the year 2000 and they have had a cancer diagnosis; then the lines will intersect (prevalence).

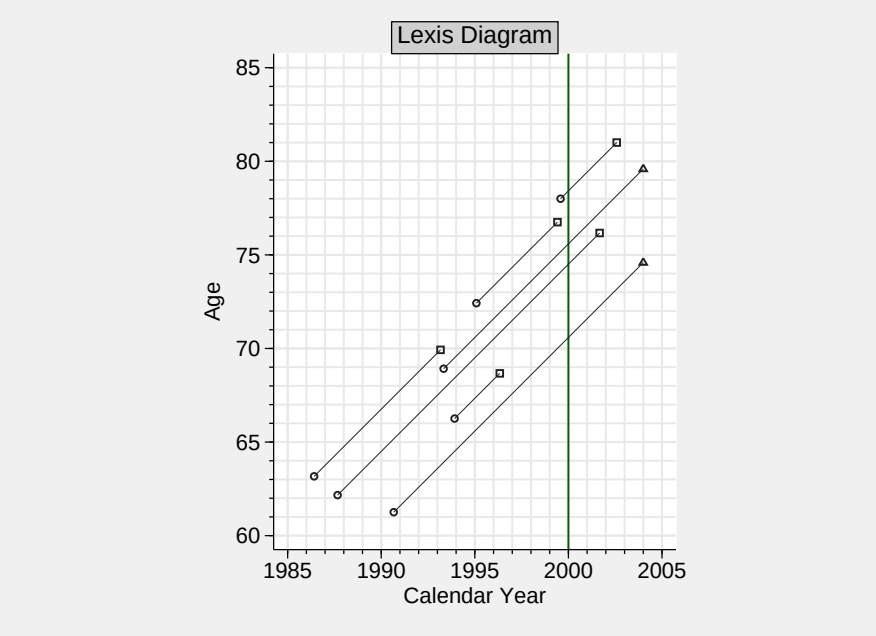

FIGURE 9.1. Lexis diagram showing how prevalence can be calculated directly.

9.3.2. Partial prevalence

A measure that attempts to account for the fact that some patients in the prevalence proportion are no longer causing an excess burden due to their cancer diagnosis is partial prevalence. Partial prevalence sets a limited time-span to how long a patient can be assumed to be a prevalent case. For example, to estimate the 10-year partial prevalence, incident cancer cases are only considered up to 10 years prior to the estimation. Those patients that were diagnosed greater than 10 years previous to the partial prevalence estimation are assumed to no longer have a prevalent case of cancer. This ties in with the idea of statistical cure that was introduced in Chapter 6. Those patients that are no longer at an excess risk of death due to their cancer may well be the patients that we want to remove from our prevalence estimates.

The concept of partial prevalence is illustrated in Figure 9.2. After 10 years post-diagnosis the line indicating the patient's follow-up time is replaced by a dashed line indicating that they are no longer considered a prevalent case. In the year 2000, this reduces the prevalent cases from 4 to 3 in the small subset of patients (compared to Figure 9.1). Partial prevalence has been used in a number of settings to calculate cancer burden estimates [Pisani et al., 2002; Tabata et al., 2008]. Partial prevalence estimates are given for 1, 3, and 5-year intervals by

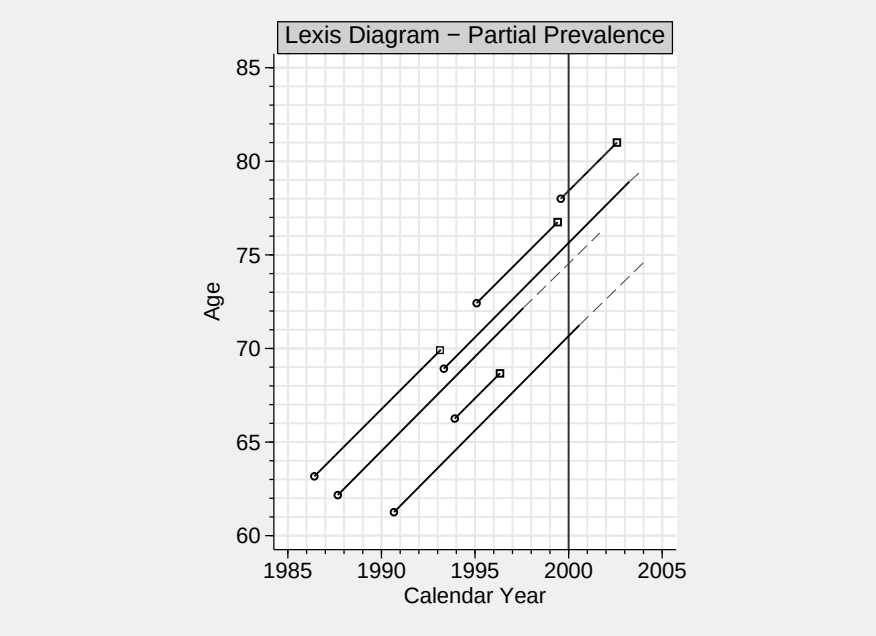

FIGURE 9.2. Lexis diagram showing how partial prevalence can be calculated directly.

Tabata *et al.* to reflect initial treatment, clinical follow-up and the point of cure over a range of cancer sites in Japan.

9.3.3. Burden prevalence

The term burden prevalence will be used to define the true measure that provides an estimate of the burden of cancer on society. That is, patients would be removed as a prevalent case if, or when, they were considered to no longer be an excess burden on society due to their cancer. This is similar in concept to the partial prevalence. However, it is not possible with the information that is available to the cancer registries to define the point at which an individual patient is no longer a burden due to their cancer. Even in the case when a patient is "cured" from the cancer of interest, it still might be the case that they require more frequent check-ups or scans than a member of the general population (of similar age and gender).

There have been a number of examples in the literature where attempts have been made to estimate something akin to this quantity [Mariotto et al., 2003; Chauvenet et al., 2009; Harding et al., 2011]. However, these approaches have used external information by linking the registry data to further follow-up information for patients. Mariotto *et al.* [2003] do this for SEER data for colon cancer patients and link to Medicare data. A similar approach is adopted by

Chauvenet *et al.* [2009] for French colorectal patients with treatment follow-up information. Harding *et al.* [2011] calculate what they term as symptom prevalence for a small sample of African cancer patients by obtaining estimates of the number of patients suffering from certain symptoms following a cancer diagnosis.

9.4. Estimating prevalence

Following a similar argument to Capocaccia and De Angelis [1997] (with differing notation), if a single birth cohort, c, is considered, let $h^*(a)$ be the general mortality in the population and $I(a)$ be the incidence rate for the disease of interest at attained age a. Further to this, let $h(\alpha_0, a)$ be the overall mortality (hazard) rate for cancer patients diagnosed at age α_0, followed up until age a. The probability of a person in the general population to be alive at attained age a in the given cohort (c), that is, in the year ($y = c + a$), is:

$$S^*(a) = \exp\left(-\int_0^a h^*(u)\,\mathrm{d}u\right). \tag{9.1}$$

The probability of being alive having had a previous diagnosis of cancer is:

$$\int_0^a I(t)\exp\left(-\int_0^{\alpha_0} h^*(u)\,\mathrm{d}u\right)\exp\left(-\int_{\alpha_0}^a h(\alpha_0, u)\,\mathrm{d}u\right)\mathrm{d}t, \tag{9.2}$$

this gives the probability of getting a cancer diagnosis (incidence), having survived until the point of diagnosis (α_0), multiplied by the probability of surviving until age a given that the patient was diagnosed with cancer at age α_0.

Therefore, under Bayes' Theorem, the proportion of the birth cohort with a cancer diagnosis at age a (prevalence, $P(y, a)$) is given by the ratio of the two previous equations:

$$P(y,a) = \frac{\int_0^a I(t)\exp\left(-\int_0^{\alpha_0} h^*(u)\,\mathrm{d}u\right)\exp\left(-\int_{\alpha_0}^a h(\alpha_0,u)\,\mathrm{d}u\right)\mathrm{d}t}{\exp\left(-\int_0^a h^*(u)\,\mathrm{d}u\right)}. \tag{9.3}$$

which simplifies to:

$$P(y,a) = \int_0^a I(t)\exp\left(+\int_{\alpha_0}^a h^*(u)\,\mathrm{d}u\right)\exp\left(-\int_{\alpha_0}^a h(\alpha_0,u)\,\mathrm{d}u\right)\mathrm{d}t. \tag{9.4}$$

Using equation 7.2, the equation can be given in terms of excess mortality, $\lambda(\alpha_0, a)$:

$$P(y,a) = \int_0^a I(t)\exp\left(-\int_{\alpha_0}^a \lambda(\alpha_0,u)\mathrm{d}u\right)\mathrm{d}t. \tag{9.5}$$

This equation gives an estimate of prevalence using a continuous representation of incidence and survival through the use of integration. An approximation to this estimate can be given through calculating estimates for given ages and calendar years and summing over the age range and calendar time.

9.5. Literature Review

Keiding [1991] gives the details of calculating total prevalence from a probabilistic perspective with examples of the calculation of prevalence and incidence in a variety of settings. He gives the key formula that will be used in this chapter and a clear exposition of the inter-relations of incidence, survival, mortality and prevalence.

Approaches to estimating the variance of cancer patient prevalence have also been considered [Gigli et al., 2006]. The outlined approach gives details of the necessary calculations to calculate the variance estimates for the complete prevalence. That is, also including the cases prior to the setting up of the cancer registry. The variance estimates are obtained via the delta method having fitted models for incidence that account for the patients age, and mixture cure models for relative survival.

Total prevalence estimates have been given using Connecticut Tumor Registry data [Feldman et al., 1986]. However, in order to give an estimate of total prevalence, the cancer registry needs to be have been set-up for a sufficient number of years in order to capture all of the necessary incidence data. The majority of the other state registries in the US do not have sufficient follow-up information to directly estimate total prevalence. An approach has been suggested in order to calculate the completeness of a given registry in terms of capturing the total prevalence [Capocaccia and De Angelis, 1997].

The approach set out by Capocaccia and De Angelis [1997] is an attempt to account for patients that were diagnosed prior to the cancer registry being set-up. This is even more of an issue if total prevalence is to be estimated. Total prevalence includes long-term survivors and, for some cancer sites, there will be a significant group of patients who were diagnosed prior to the registry being set-up who will therefore not be in any of the incidence calculations. Using the information in the observation period, it is possible to give an estimate of the number of patients diagnosed prior to the inauguration of the registry [Capocaccia and De Angelis, 1997].

An extension of this approach has been used in order to estimate the number of people who have had a previous childhood cancer diagnosis has also been developed [Simonetti et al., 2008].

Using mixture cure models, it is possible to split the patients into different categories according to their survival experience [Coldman et al., 1992]. This approach has been used for colon cancer patients to ascertain whether patients experience similar mortality to the general population, or whether the cancer diagnosis leads to a premature death [Gatta et al., 2004]. The authors argue that this helps further subdivide the prevalent cancer patients into the type of care that will be required.

A similar approach using mixture cure models has been applied and the potential to project the future cancer prevalence has been considered [Phillips et al., 2002]. This approach again attempts to incorporate statistical cure into the prevalence estimates in order to account for the fact that a proportion of cancer patients do not experience an excess mortality due to their cancer diagnosis.

There have been numerous examples of the methods for estimating prevalence being applied for cancer registry data. National cancer prevalence estimates have been given for France [Colonna et al., 2000] using relative survival and age-cohort models for incidence. Merril *et al.* [2000] calculated prevalence for SEER data using incidence and relative survival cure models. They adopted the approach to calculate the completeness index [Capocaccia and De Angelis, 1997] in order to obtain estimates of complete prevalence.

Cancer prevalence estimates have also been given for the United Kingdom [Maddams et al., 2009] with projections based on trends in past prevalence rather than fitting separate models for incidence and survival. Partial prevalence estimates were calculated, and were referred to by the authors as limited-duration prevalence counts. These estimates were then projected backwards in order to estimate the complete prevalence estimates for each cancer site.

Health and planning officials need to plan future treatment and care for the population. It is therefore of interest and importance that accurate projections of the future cancer burden can be obtained from the currently available data [Theisen, 2003]. The main projection approaches for estimating prevalence involve projecting estimates of incidence and survival [Verdecchia et al., 2002; Heinävaara and Hakulinen, 2006]. These will be given further attention in the following section.

There have been a number of examples that also include projections of prevalence. There have been projections using kidney cancer data in Italy [Bosetti et al., 2009], where two approaches to projection have been compared. The authors compared the future prevalence assuming that the rates remained the same at the end of the observed data, to the future prevalence obtained assuming that there is a 1% annual decrease over the projection period. The idea of using two hypothesis for the projections in order to give a range of potential values for prevalence is one that has been commonly adopted [Capocaccia et al., 1997; Verdecchia et al., 2002].

Some authors have used the term cancer burden to refer to the total number of new cases in the future [Coupland et al., 2010]. Calculating the incidence and survival separately allows the exploration of the trends in terms of both of these key quantities alongside the combined trends that is then implied for the prevalence of the disease. The method of projection developed in Chapter 4, can be used to estimate the future total number of cases when combined with projected population figures. A combination of incidence and mortality rates have also been used to define the cancer burden [Arbyn et al., 2007; Chan et al., 2004]. Arbyn *et al.* [2007] look at the Europe-wide burden of cervical cancer by assessing the number of new cases, and the number of deaths relating to the disease for the year 2004. Chan *et al.* [2004] adopt a similar approach using SEER data for prostate cancer. They also make projections of the future mortality associated with prostate cancer in the United States with the prospect of an ageing population.

An approach using the Joinpoint software [Kim et al., 2000] and GLOBOCAN software [Ferlay et al., 2010] to make incidence and mortality projections has also been applied to data from Saudi Arabia [Ibrahim et al., 2008]. The future breast cancer burden in Saudi Arabia is considered by projecting the current rates into the future.

A number of global estimates for cancer burden have been given [Parkin et al., 2001; Parkin, 2001; Pisani et al., 2002; Ploeg et al., 2009]. Parkin [2001] reports the global incidence, mortality and 5-year partial prevalence estimates for the year 2000. Pisani *et al.* [2002] give partial prevalence estimates of 1 year, 2-3 year, and 4-5 years, and argue that these relate to initial treatment, clinical follow-up, and point of cure respectively for the majority of cancer sites. They employ an approach of adding up the total number of cases, multiplied by the

relevant overall survival proportions in order to estimate the partial-prevalence. They also give combined country estimates of prevalence for the most prevalent cancer sites, and report the estimates for the developed, and developing countries separately. Ploeg *et al.* [2009] concentrate on bladder cancer, but give world-wide estimates of incidence and mortality for that site. They argue that global changes in risk factors will lead to major increases in the incidence and prevalence of the disease in the future.

The methods have also been applied to other diseases, such as coronary heart disease [Tobias et al., 2008]. Provided that the population-based data is obtainable, it is possible to provide estimates of incidence, mortality, survival and prevalence for any disease of interest.

Recently, methods to try to ascertain the proportion of cancer cases that were associated with radiotherapy treatment for an earlier cancer have been applied to UK data [Maddams et al., 2011]. This gives an interesting further insight into the potential burden of cancer on the basis that second malignancies could be caused by treatment for an earlier form of cancer.

9.6. Combining Incidence and Survival estimates

9.6.1. Introduction

A number of techniques have been proposed that combine the results of the survival and incidence estimations to give an estimate of prevalence. The relationship between prevalence, survival and incidence is clearly defined, and can be reduced to a simple sum of the multiplied survival and incidence estimates provided that the data from the appropriate time periods are used.

9.6.2. Current Methods

9.6.2.1. *Heinävarra and Hakulinen Method*

One method that has been proposed to obtain model-based prevalence estimates combines the overall survival with the number of cases within each calendar year and age group [Heinävaara and Hakulinen, 2006]. This is performed by the summation of the appropriate log-likelihoods, and uses individual level data. The paper also gives details of the potential to predict these estimates into the future. However, confidence intervals are not given for the predicted estimates. The models that are used for the incidence data are age-cohort models with an included drift

term. The survival models are based on mixture cure models, but the overall survival is used when evaluating the prevalence estimates.

Rather than using the probabalistic argument given in the previous section, the approach laid out by Heinävarra and Hakulinen [2006] uses a more standard approach to estimating prevalence with the total number of cases, and the overall survival.

The estimate of $P(y,a)$ is given by:

$$\frac{\left\{0.5A(y,a)\int_0^{0.5} S(0.5-u;y,a)\mathrm{d}u + \sum_{l=1}^{v-1}\left\{A(y-l,a-l)\int_0^1 S(l+0.5-u;y-l,a-l)\mathrm{d}u\right\}\right\}}{N_{y,a}}, \quad (9.6)$$

where $A(y,a)$ is the total number of new cases of cancer in the population in year y and age a, and $N(y,a)$ is the total number of people in the entire population as a whole of age a in year y. In this case, the survival function $S(t;y,a)$ relates to the overall survival. In the method outlined in Section 9.4, the relative survival function is multiplied by the incidence rate to achieve the prevalence estimate. The approach given in Equation (9.6) sums up prevalence estimates over yearly intervals to obtain the partial-prevalence estimates defined by the v-year length of follow-up. In the approach set out in Equation (9.6), the calculation is performed in terms of the number of new cancer patients (which accounts for the population structure in the cohort) and therefore overall survival is required to obtain the prevalence estimates.

9.6.2.2. *PIAMOD Method*

The PIAMOD (Prevalence, Incidence, Analysis MODel) method is a two-step procedure to estimate prevalence. The first stage is to fit age-period-cohort models to the data, and use those models to project incidence into the future. The second stage involves combining those estimates with the modelled survival results [Verdecchia et al., 2002]. The methods for obtaining projections for the survival values involved simply fitting two cases; assuming the survival remained constant (conservative), and assuming that the rate of improvement remained the same as in the observation period (optimistic). Software is available to carry out the PIAMOD method [Verdecchia et al., 2002].

The PIAMOD approach is based on the equations given in Section 9.4, and uses relative survival estimates for the survival component. The approach has been used in practice to obtain estimates of future prevalence [Mariotto et al., 2011; Yu et al., 2011]. Mariotto *et al.* [2011] use the PIAMOD approach to obtain future estimates of prevalence for SEER data prior to

multiplying those estimates by annualised net costs to calculate a value of financial burden. Yu *et al.* [2011] provide a protocol for a study to be undertaken using Australian prostate cancer patients. They intend to use the PIAMOD approach to obtain future estimates of prevalence by stage of care and link to other data sources to obtain information of quality of life measures for prostate cancer patients.

9.6.2.3. *Own Work*

Projection estimates can be given using the incidence projection methods described in Chapter 4 and the survival models that are described in Chapter 8. Using restricted cubic splines for both the incidence and survival components, and ensuring the data is split finely, means that continuous estimates of prevalence can be estimated. Further to this, age-specific estimates of the prevalence of the disease can be estimated, as well as the age-specific incidence and survival contributions to that estimate. The previous approaches have failed to fully account for the effects of age, period and cohort when calculating and projecting incidence. The approach outlined in Chapter 4 does account for each of those effects and the newly proposed incidence projection technique appears to perform better than the previous approaches. Accounting for the effect of cohort allows the effect of period to vary for different values of age. This is similar in principle to the interaction that was fitted between age and year of diagnosis for the survival models described in Equation (8.8).

9.7. Projecting Prevalence vs Combining Incidence and Survival

Prevalence estimates can be obtained from the registry data directly by simply counting the proportion of patients that are still alive having had a diagnosis of cancer. Many of the methods for estimating prevalence instead adopt an approach to modelling incidence and survival in order to have a model-based estimate of prevalence. In the following sections, approaches that combine the models for incidence and survival will be given. The model-based estimates of incidence and survival have been verified in the previous chapters, and methods to project these estimates have been developed and improved to give realistic projections from the available data. Projecting the two quantities separately prior to combining them to obtain projected prevalence allows a greater control and understanding over the trends that are being projected. Simply projecting the counting-based prevalence estimates may not account for how the incidence and survival trends have combined to give the prevalence. In addition to this, the projected incidence

estimates can be used to calculate the number of new cases in the future, which is an important quantity in itself when trying to gauge the cancer burden on society. An example using the UK data adopting an approach of projecting the prevalence estimates (rather than separate incidence and survival estimates) has been used to obtain short-term projected estimates of prevalence [Maddams et al., 2009].

9.8. Description of Methods for Model-based Prevalence Estimation

The results for the prevalence estimation are given for two key example cancer sites; lung cancer and female breast cancer. Model-based prevalence estimates from the combination of the incidence and survival models are given within the range of the available data so that comparisons can be made to the true prevalence estimates obtained by the counting method. For the estimates within the range of the data, the models described in Chapter 2 are used for the incidence models. There is no need to consider moving the boundary knot within the range of the data until the projections are being made. For the relative survival models, modelled cohort analysis techniques are employed as described in Chapter 8. A linear effect of year of diagnosis is fitted, and the effect of age is described by a spline function with 10 degrees of freedom. Such a high degree of freedom was selected to try to fully capture the shapes in the real data. It is unlikely that such a high degree of freedom will be necessary in practice. A similar approach was adopted for the spline terms for the APC model, with 15 degrees of freedom selected for each of the age, period and cohort. The results of Chapter 3 show that this degree of complexity is rarely required.

The estimates of prevalence are obtained by summation as has been described by Verdecchia *et al.* [2002]. The approach is based on the equations that are given in Section 9.4. The X-year partial estimates are defined by only considering incident cases that have occured within the last X years when calculating the prevalence estimates. The prevalence estimates are compared to monthly estimates that can be obtained by a counting approach using the relevant data from the cancer registry data. The counting approach is simply a count of the number of patients that are still alive in the relevant month, that have had a diagnosis of cancer within the last X years. These counting based estimates should be similar to the estimates obtained from combining the model-based survival and incidence estimates.

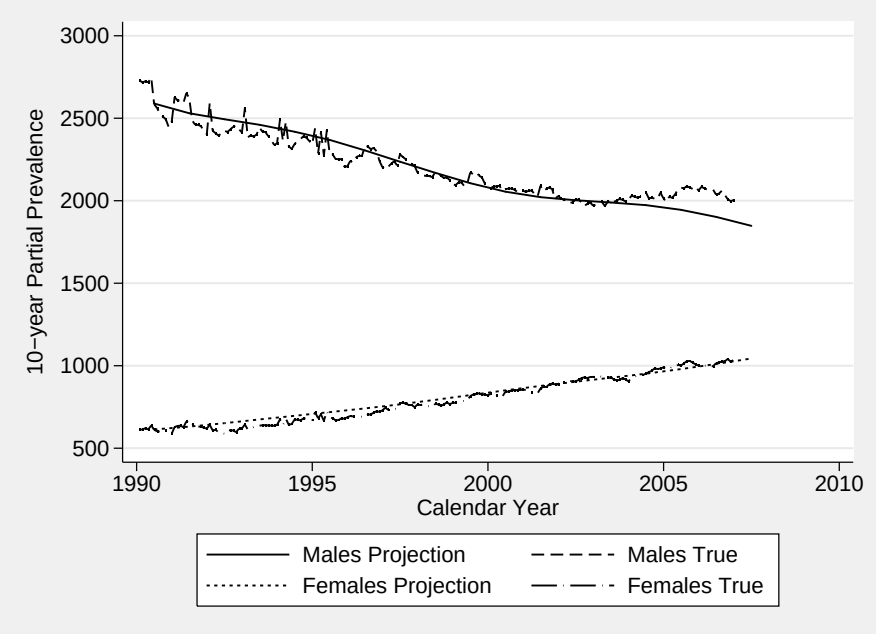

FIGURE 9.3. Total (over all ages) partial (10 year) prevalence for lung cancer separated by gender with truth.

Projected estimates are given by combining the projected estimates of incidence and survival according to the methods that were outlined in Chapters 4 and 8 respectively. A sensitivity analysis to the levels of the components that can be varied when making the projections will be undertaken.

9.9. Results

9.9.1. Lung Cancer

Figure 9.3 shows the model-based estimates of the total number of prevalent cases (10-year partial prevalence) for lung cancer over calendar year for males and females. It is clear that the prevalence is decreasing for males, whereas it is increasing for females. This is likely to be directly related to the current smoking trends for the Finnish population. The shape of the curves are similar to the shape for the incidence curves (see Figure 9.5). The prevalence and incidence estimates are more similar for lung cancer than for other sites due to the poor long-term survival associated with a diagnosis of lung cancer (see Figures 9.7 and 9.8). The graph also compares the true partial prevalence estimates that are obtained from the counting method in monthly intervals. The model-based estimates seem to perform well for females, but

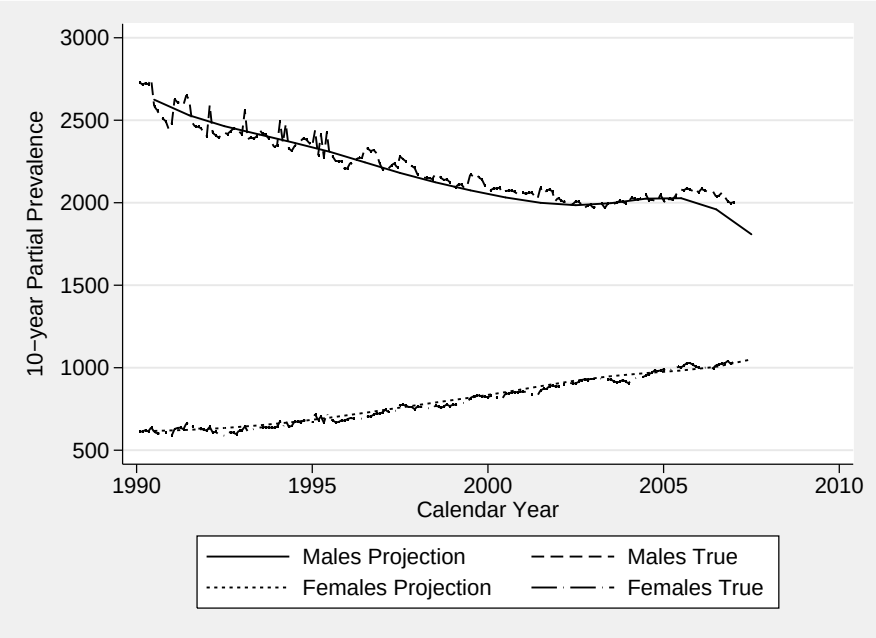

FIGURE 9.4. Total (over all ages) partial (10 year) prevalence for lung cancer separated by gender with truth using splines instead of linear year.

provide an underestimate for males after 2003. The model-based estimates are a combination of the modelled effect of incidence and survival. For the survival models, a linear effect of year of diagnosis was assumed over a long range of years and this results in the periods of underestimation. This is similar to the issues that were uncovered for lung cancer in the previous chapter which were highlighted in the comparison of Figures 8.10 and 8.11. In order to counteract this, spline terms can be fitted for the effect of year of diagnosis to highlight how this can improve the fit. The results of fitting a simple spline function (4 degrees of freedom) for year of diagnosis can be seen in Figure 9.4. This results in an improved fit for both males and females, and increasing the complexity of the spline function will result in a further improvement to the fit. However, using a spline function to provide an improved fit to the current data is likely to result in less stable projections when extending the modelling approaches to project prevalence.

Figure 9.5 gives the total number of new cases (incidence rate multiplied by population size) over the same range of years for lung cancer. This is essentially the summation over age of the terms that are modelled fully by age, period and cohort for the incidence models. Again, the true total number of cases can be calculated by a process of counting, and a comparison

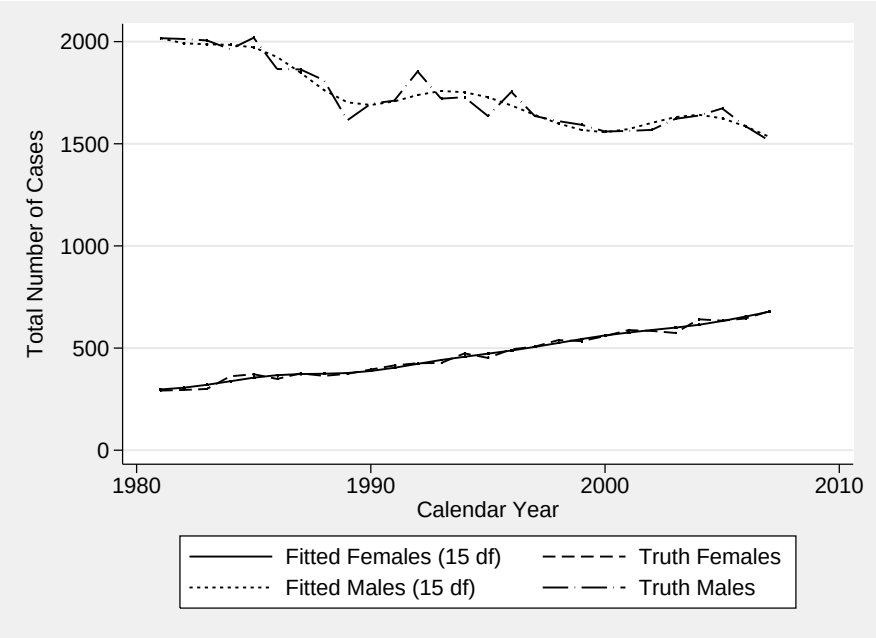

FIGURE 9.5. Total (over all ages) number of cases for lung cancer with truth.

is given to a model that has 15 degrees of freedom for age, period and cohort. The incidence models appear to fit very closely to the truth, and using a high degrees of freedom for each of the terms ensures a close fit to the true data.

Figure 9.6 gives a comparison of the fit of two different age-period-cohort models; one with 15 degrees of freedom for each of the components and one with 5. The true shape is captured much more closely by increasing the degrees of freedom for the incidence model. In the case of using 15 degrees of freedom, there is clearly a case of over-modelling in that local departures are being captured by the splines. Fully capturing the local deviations in the available data is unlikely to lead to stable projections. This is the main reason for the suggestion of moving the boundary knot within the range of the known data in order to stabilise the projections; the topic of Chapter 4.

Partial prevalence has been calculated for lung cancer using a 10 year window. The results in Figure 9.7 show the survival experience for the patients that make up the total prevalence (at least considering cases diagnosed after 1953; the start of the registry) for the year 2000. The histogram shows the proportion of patients who have a given length of time since diagnosis in yearly intervals. Around 30% of the prevalent lung cancer cases in 2000 were diagnosed

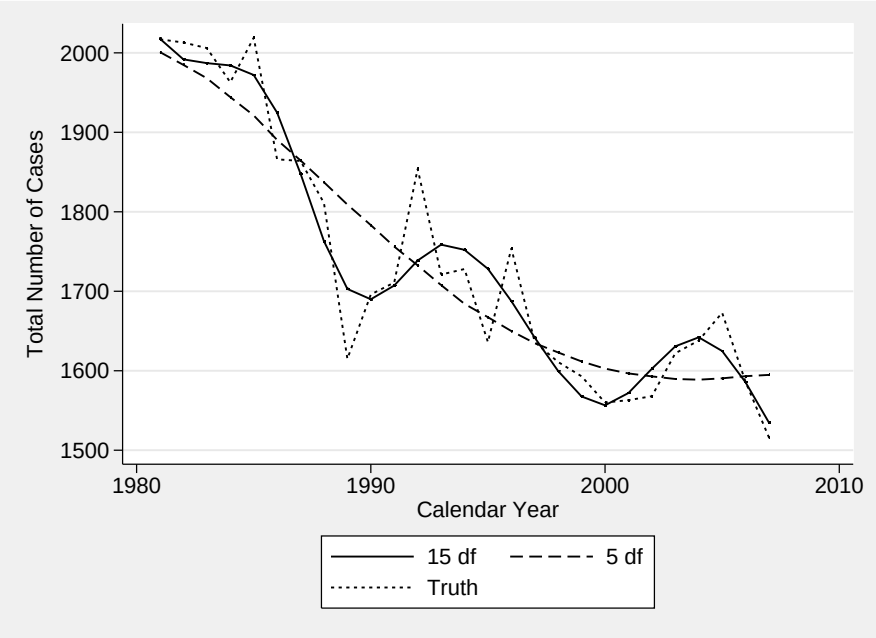

FIGURE 9.6. Total (over all ages) Number of cases for Lung cancer with truth for males.

under 1 year ago, and roughly 78% of patients would have been included in the 10-year partial prevalence estimates. It is also clear that a small proportion of patients were diagnosed in excess of 40 years prior to being classed as a prevalent case in the year 2000. Cure models can be used to attempt to assess whether or not these patients can reasonably be assumed to be cured of their disease.

Figure 9.8 gives a comparison of the relative survival experience for lung cancer patients in a 55-64 age-group. The two lines indicate two flexible parametric modelling approaches; one assumes statistical cure after 10 years by fitting a constrained model (as described in Section 6.8.5) whereas the other line is produced by an unconstrained flexible parametric relative survival model. The cure assumption forces a plateau at 10 years for the relative survival curve. The unconstrained model also appears to be flattening out, and the assumption of cure after 10 years does not appear to be entirely unreasonable. It should be noted that the cure proportion is a little over 15%, highlighting that a high proportion of lung cancer patients are likely to die quickly following diagnosis. This explains the similarities between the incidence and prevalence estimates given in Figures 9.4 and 9.5.

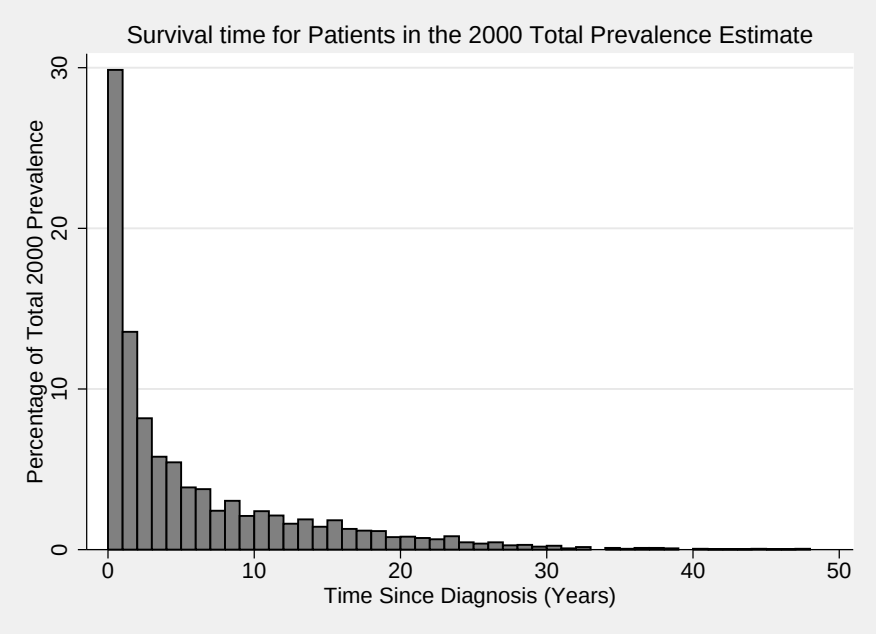

FIGURE 9.7. Time since diagnosis for lung cancer in 2000 given as the percentage of total prevalence.

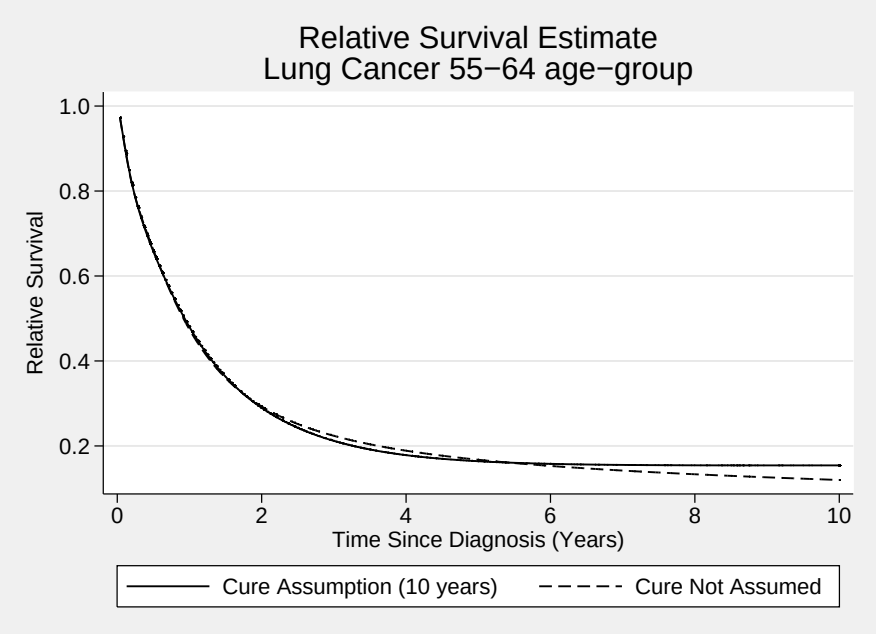

FIGURE 9.8. Evaluating the assumption of cure at 10 years for lung cancer.

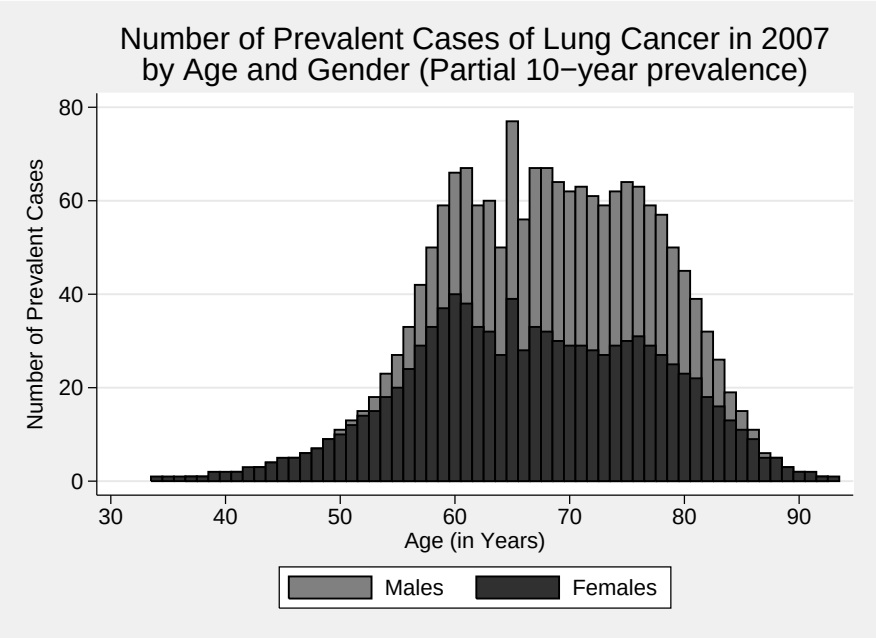

FIGURE 9.9. Age-specific partial (10-year) prevalence for lung cancer separated by gender. Females overlaid over males.

An advantage of modelling age continuously using splines is that age-specific estimates can also be given alongside the estimates over all ages. Figure 9.9 shows the model-based age-specific estimates of the total number of (partial) prevalent cases in 2007 for each of the genders. It is clear that there are still a greater number of prevalent cases for males than females over the majority of the age-range in spite of the fact that the incidence is on the increase for females, whilst it is decreasing for males. If the current smoking trends persist, then it is likely that these figures will get increasingly similar as time progresses.

The prevalence estimates are obtained from the combination of model-based estimates of incidence and survival. Therefore, it is also possible to report age-specific summaries of incidence and survival by gender. Figure 9.10 shows the number of new cases in 2007 for lung cancer. It is clear that there is a similarity between the estimates in this graph and those contained in Figure 9.9. This is due to the fact that the relative survival from lung cancer is low, particularly for the ages that have the highest number of new cases. This information can be summarised by reporting the age-specific 5-year relative survival over the age range for both genders; given in Figure 9.11. The relative survival estimates are higher for females than males which explains why the prevalence estimates are closer together for the genders than they were

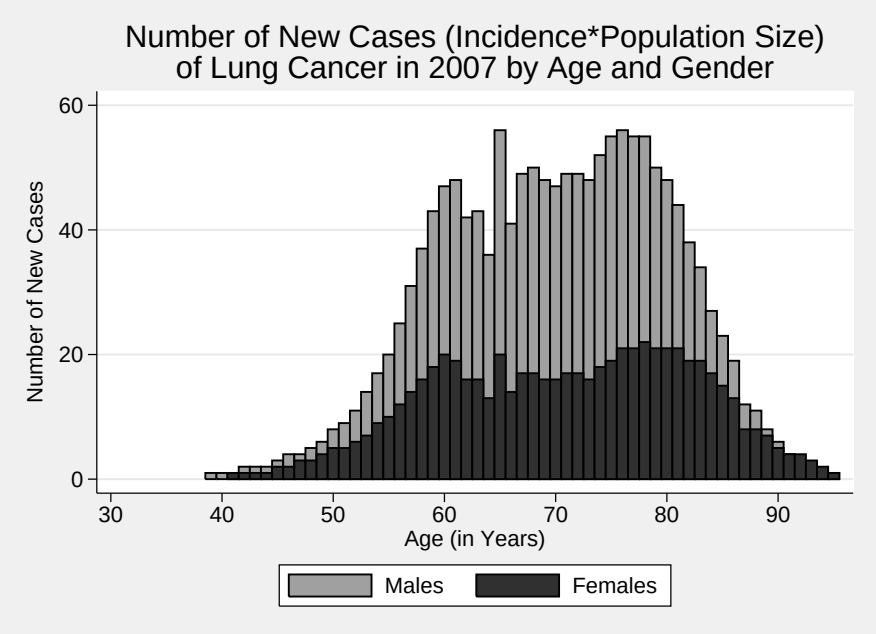

FIGURE 9.10. Age-specific total number of new cases for lung cancer separated by gender. Females overlaid over males.

for the incidence estimates; a higher proportion of female patients tend to live longer with their diagnosis of lung cancer.

One assumption that is often made as a "lower bound" (conservative estimate) when projecting prevalence is to assume that the survival rate remains the same in the future as at the last observed data point [Verdecchia et al., 2002; Heinävaara and Hakulinen, 2006]. The implications of this assumption can be assessed by using a comparison between two different scenarios. Figure 9.12 shows the partial prevalence estimates under two scenarios. The first scenario assumes that the survival experience for the patients remain the same as in 1987 for the following 20 years of follow-up. The second scenario shows what happens when the survival experience is modelled with a spline term for year of diagnosis. Both of the scenarios have the same incidence pattern from 1987 onwards. The difference for lung cancer between the two scenarios is not pronounced. This is due to the fact that there has been relatively little improvement in relative survival for lung cancer over the time-period. This is particularly clear for lung cancer for females. In the case of lung cancer, making an assumption that the relative survival stays the same as the last observed value is not as unreasonable as for other cancer sites.

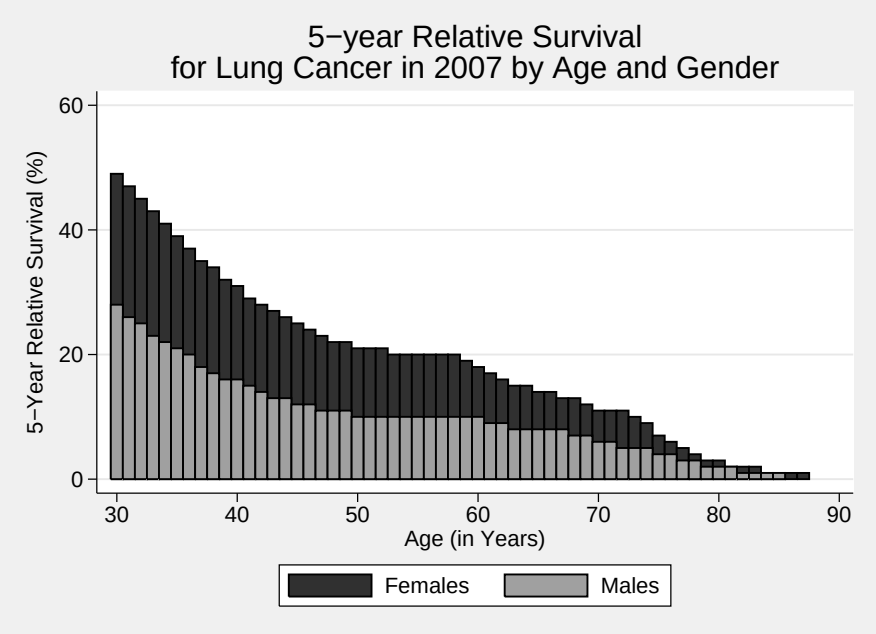

FIGURE 9.11. Age-specific 5-year relative survival for lung cancer separated by gender. Males overlaid over females.

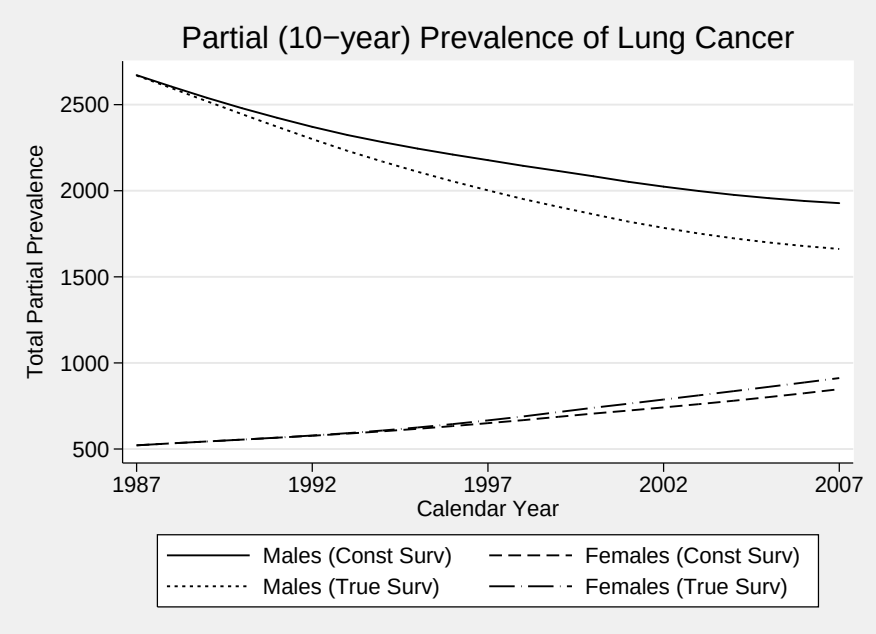

FIGURE 9.12. Checking the constant survival assumption for lung cancer separated by gender. Constant survival assumed from 1987.

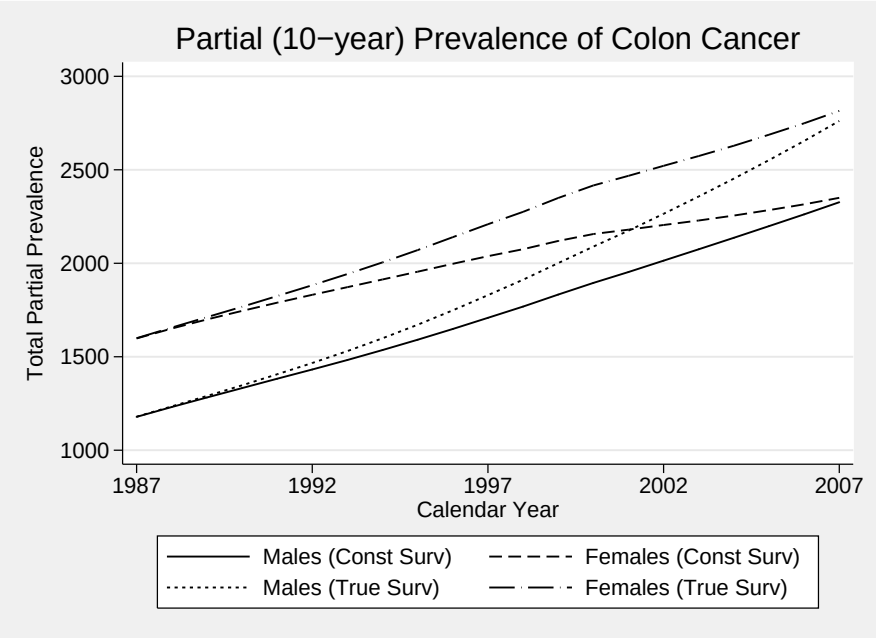

FIGURE 9.13. Checking the constant survival assumption for colon cancer separated by gender. Constant survival assumed from 1987.

Figure 9.13 shows the partial prevalence estimates under the same two scenarios for colon cancer as a comparison. Colon cancer survival has been shown to be improving over time in Finland [Teppo et al., 1999]. It is clear that this can lead to a significant underestimation of the total prevalence if patient survival continues to improve over the projection period. It seems as though both genders have a similar improvement in their survival over the time-period for colon cancer. The differences when this analysis was conducted for lung cancer were not as pronouced. This is because colon cancer survival is improving over calendar time whereas less of an improvement is seen for lung cancer survival over the same time period.

Lung cancer provides an example with poor relative survival meaning that the prevalence estimates are not too dissimilar to the incidence estimates. Modelling survival over a long range of data and assuming a linear effect for year of diagnosis can result in the model-based prevalence estimates not truly fitting the observed prevalence. Therefore, when making projections, a key decision lies in how long a data range to use in order to estimate the linear effect that is to be projected. Also, when making long-term projections of prevalence, it is unlikely that the projected linear effects for both survival and incidence will still hold. There is increasing uncertainty in the projection assumptions as time since the end of the observed data increases.

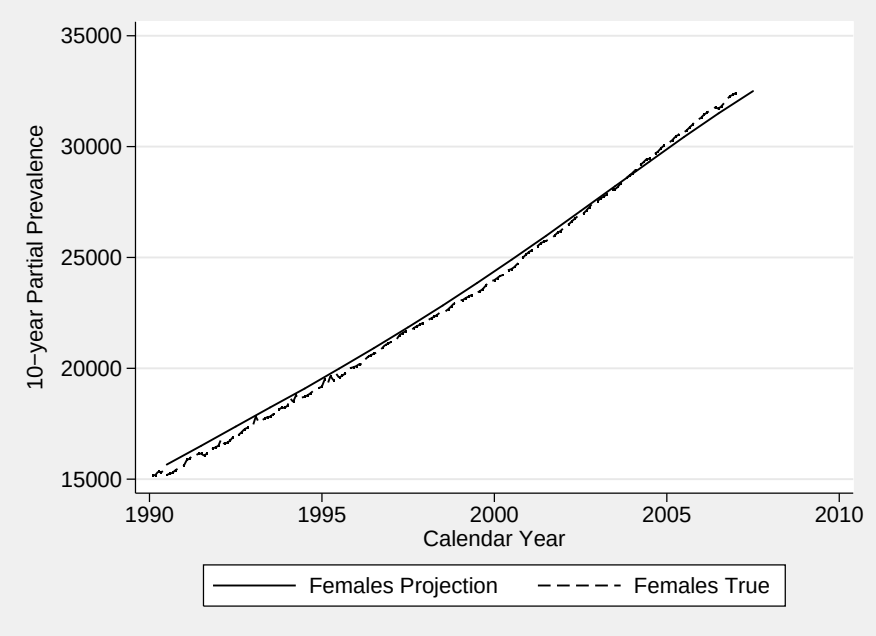

FIGURE 9.14. Total (over all ages) partial (10 year) prevalence for breast cancer for females With Truth.

9.9.2. Breast Cancer

Figure 9.14 compares the model-based 10-year partial prevalence estimates for breast cancer to the observed monthly-split counts for prevalence. An interaction between the spline terms for age and the linear effect of year of diagnosis was fitted in the survival model in light of the results seen in the previous chapter (see Figure 8.12). It is clear from the figure that breast cancer prevalence is increasing over calendar time. The absolute numbers are also much larger than those seen for lung cancer. The fit to the true data is fairly good; the deviations are likely to be due to the fact that a linear trend for year of diagnosis is assumed over an extended period.

Figure 9.15 shows the corresponding fit of the incidence model that was applied to obtain the prevalence estimates given in Figure 9.14. In the case of breast cancer, the number of new cases is far fewer than the number of 10-year partial prevalence cases in each calendar year. This is because the relative survival for breast cancer is far higher than that of lung cancer, and therefore, the prevalence estimates also include a higher proportion of cases that were diagnosed up to 10 years previously. The fact that the underlying shape is quite simple and that there is

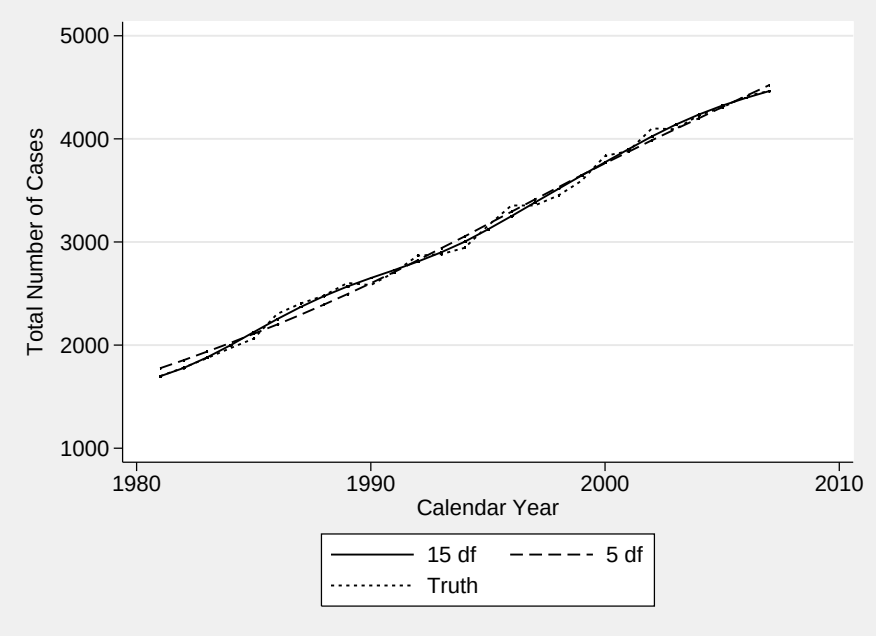

FIGURE 9.15. Total (over all ages) Number of new cases for Breast cancer Females With Truth.

a large amount of information results in the fit of the models with 5 and 15 degrees of freedom being fairly similar (see Chapter 3 for more details on the number of knots for APC analyses).

Figure 9.16 shows the survival experience for the patients that make up the total breast cancer prevalence (at least considering cases diagnosed after 1953; the start of the registry) for the year 2000. There is quite a spread across the number of years since diagnosis for breast cancer. There is a larger proportion (around 31%) that are diagnosed greater than 10 years ago than there was for lung cancer (around 22%). However, the important factor is whether or not their cancer diagnosis means that they are still a burden on society or not. One way to evaluate this is to assess whether or not the patients still have an excess mortality associated with their cancer diagnosis X years after diagnosis. This can be assessed through cure models (see Section 6.8).

Figure 9.17 compares the assumption of cure at 10 years to an unconstrained model. It is clear from the comparison of the two lines that forcing a plateau after 10 years for breast cancer is entirely unreasonable. When calculating the 10-year partial prevalence, cases of breast cancer diagnosed greater than 10 years ago are not included in the estimate. However, it is clear from the comparison in Figure 9.17 that breast cancer patients are still at an excess risk

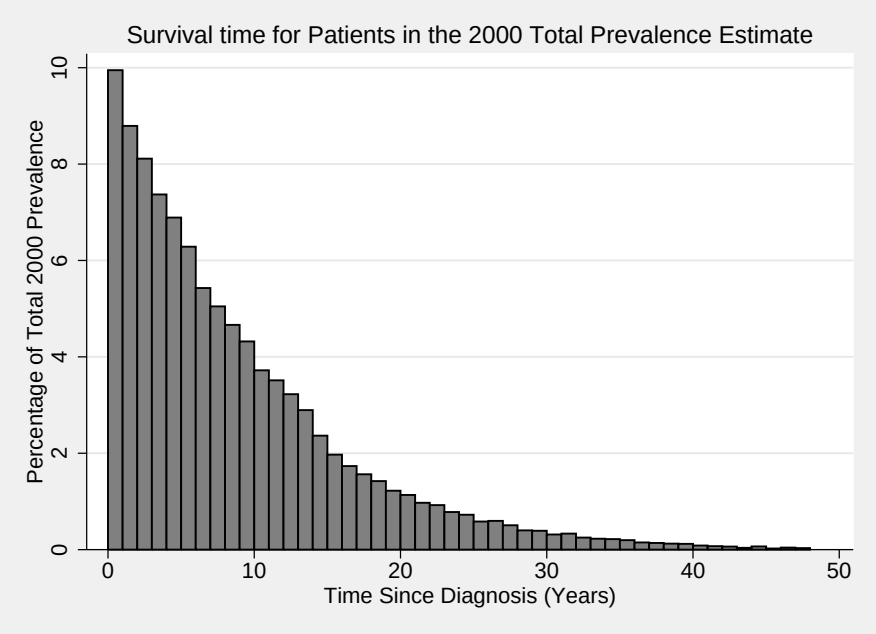

FIGURE 9.16. Time since diagnosis for breast cancer in 2000 given as the percentage of total prevalence.

of death due to their cancer diagnosis (when compared to the general population). As was mentioned in Section 6.8.6, women who are diagnosed with a case of breast cancer are at an increased risk of death compared to the general population for many years, and it has been argued that a cure point is not actually reached even after 20 years from the diagnosis [Brenner and Hakulinen, 2004; Woods et al., 2009]. Consequently, it is possible to give a longer-range partial projection estimate for breast cancer to try to incorporate the potentially burdensome cases that are excluded.

Figure 9.18 instead compares the model-based 20-year partial prevalence estimates for breast cancer to the true prevalence data. The numbers in the figure can be compared to those obtained for the 10-year partial prevalence estimates. In 1990, the two estimates are approximately 5000 patients apart whereas in 2007 the difference is closer to 13,000 patients. This highlights that there has been an improvement in terms of long-term breast cancer survival over time. Considering that there is an increased excess mortality over 20 years after a breast cancer diagnosis, it could be argued that the 20-year partial prevalence is a closer estimate to the burden prevalence for breast cancer. However, most of the long-term patients will be treated for other diseases and a large proportion will not live long enough to contribute 20 years

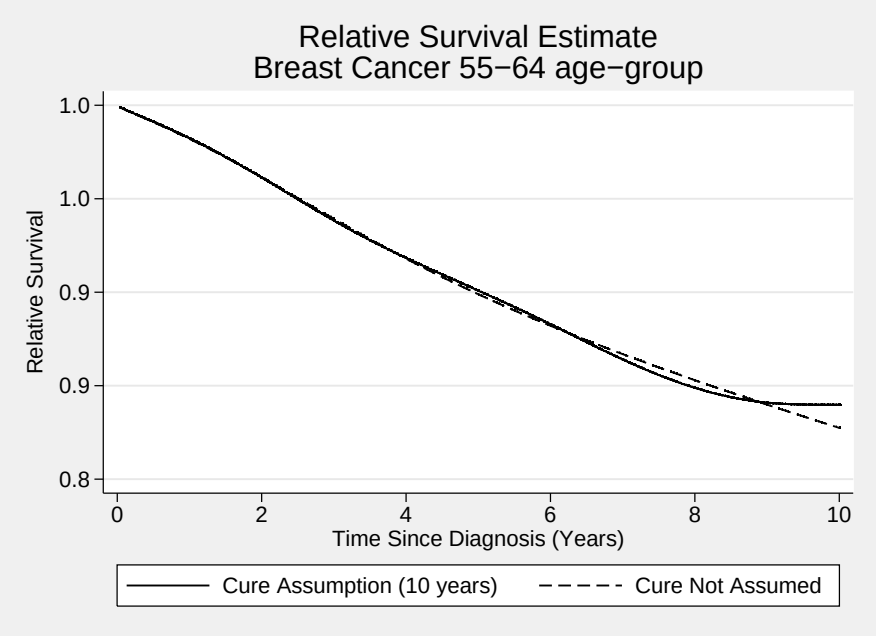

FIGURE 9.17. Evaluating the assumption of cure at 10 years for breast cancer.

of burden. It is difficult to disentangle those who are still a burden from their original cancer diagnosis without extra information on the follow-up of the patients.

Figure 9.19 gives a comparison of the 10- and 20-year partial prevalence estimates for female breast cancer. There are two comparisons made as part of the figure. The top part of the figure gives the comparison of the prevalence rates per 100,000 patients. The 20-year partial prevalence rate is consistently higher than the 10-year rate across the entire age-range. This is because the patients that were diagnosed between 10 and 20 years previously are also included in that estimate. The lower part of the figure gives the same comparison in terms of the total number of prevalent cases. This accounts for the population structure in the year 2007. The large differences observed for the older ages have less of an effect in real terms because it is less likely that the patients will make it to this age.

9.10. Projection

The future cancer burden can also be calculated by combining the projected incidence and survival estimates. Using a similar approach to the sensitivity analysis adopted in Section 4.6.1, it is possible to give the projected estimates under a range of scenarios. There are six key

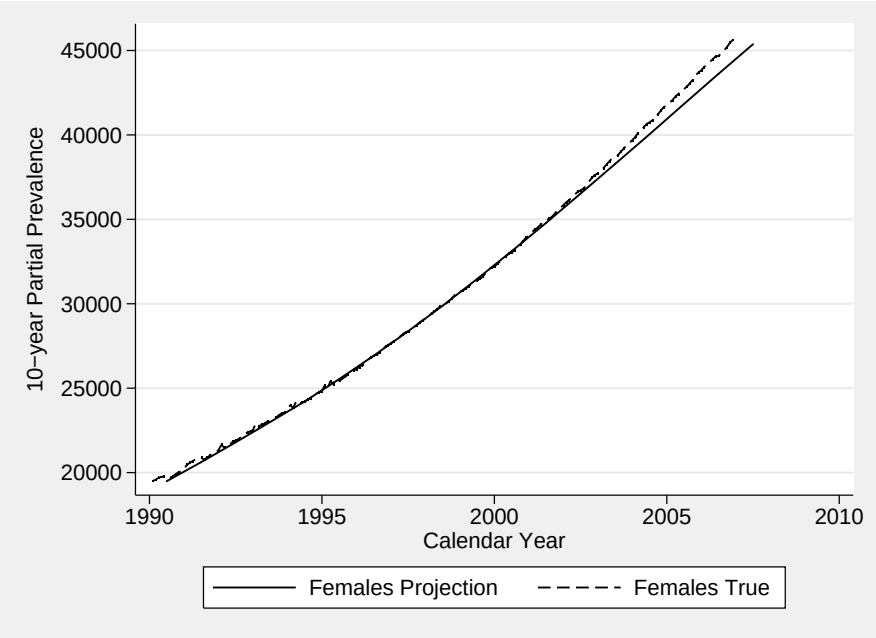

FIGURE 9.18. Total (over all ages) partial (20 year) prevalence for breast cancer for females With Truth.

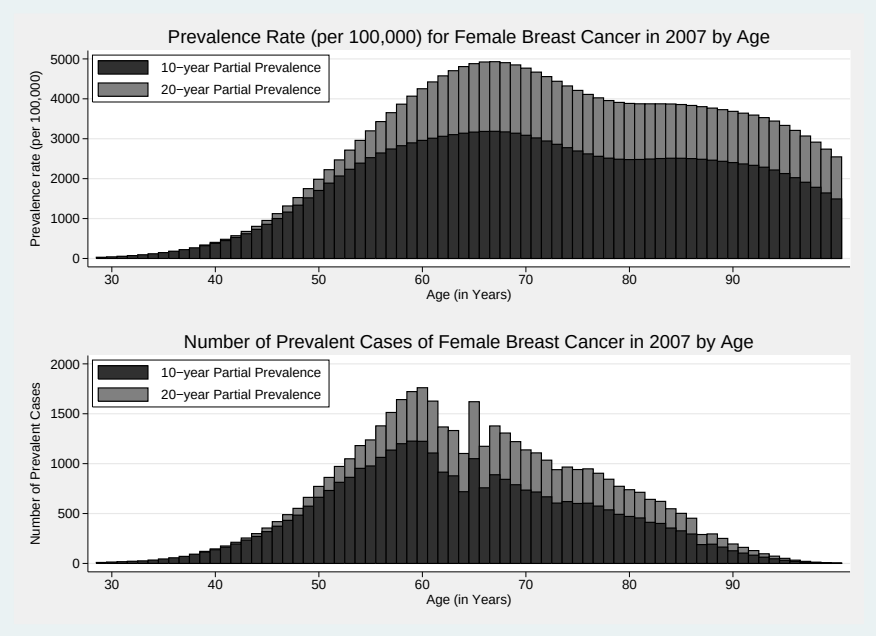

FIGURE 9.19. 10- and 20-year partial prevalence for female breast cancer in 2007. 10-year partial prevalence overlaid over 20-year partial prevalence. The top graph gives the prevalence rate per 100,000 people, whereas the bottom graph gives the actual number of prevalent cases by accounting for the population structure in 2007.

variants across both the incidence and survival projection model. The key quantities for the survival model are the number of degrees of freedom used for the spline term for age, and also the length of data (range of years of diagnosis) used to define the linear trend improvement over year of diagnosis. The four variants for the incidence models are made up of the three choices of degrees of freedom (age, period and cohort) and also the length of time that the boundary knot is moved within the range of the data (referred to as Linear Length).

	Survival		**Incidence**			
Scenario	**Start Year**	**df Age**	**Linear Length**	**df Age**	**df Period**	**df for Cohort**
1	1995	10	10	8	5	8
2	2000	10	10	8	5	8
3	1995	15	10	8	5	8
4	1995	10	10	8	8	8
5	1995	10	13	8	5	8
6	2000	10	7	8	5	8
7	1995	15	10	5	5	5
8	1998	10	10	5	5	5

TABLE 9.1. Sesitivity Analysis Scenarios

Eight scenarios were considered that took various values for the 6 variants for both the lung cancer data for males, and the breast cancer data for females. The values chosen for each scenario are summarised in Table 9.1. Scenario 1 was considered to have the most appropriate values selected for each of the variants based on the information from the previous sensitivity analyses in the earlier chapters (Chapters 4 and 8).

Figure 9.20 shows the results of the 8 scenarios for female breast cancer, with projections from the end of 2007 up until 2030. The long-term projections are likely to be less reliable considering the linear projection was made from data that was observed a long time previously. It is clear that Scenarios 2 and 6 give lower projected estimates than the other six scenarios. These two scenarios have a common variant for the survival models in that they both have a start year of 2000 for the survival model. It appears that using a shorter range of data for the survival model leads to different projected estimates; the linear trend over the more recent data takes a lower value than those that start at 1995. Looking at the estimates for the "true" age-standardised relative survival for breast cancer in Figure 8.3, it is clear that there is a decrease in the gradient of the line beyond the year 2000. This is the reason why the projected estimates are lower under Scenarios 2 and 6, and to a lesser extent under Scenario 8 (start year

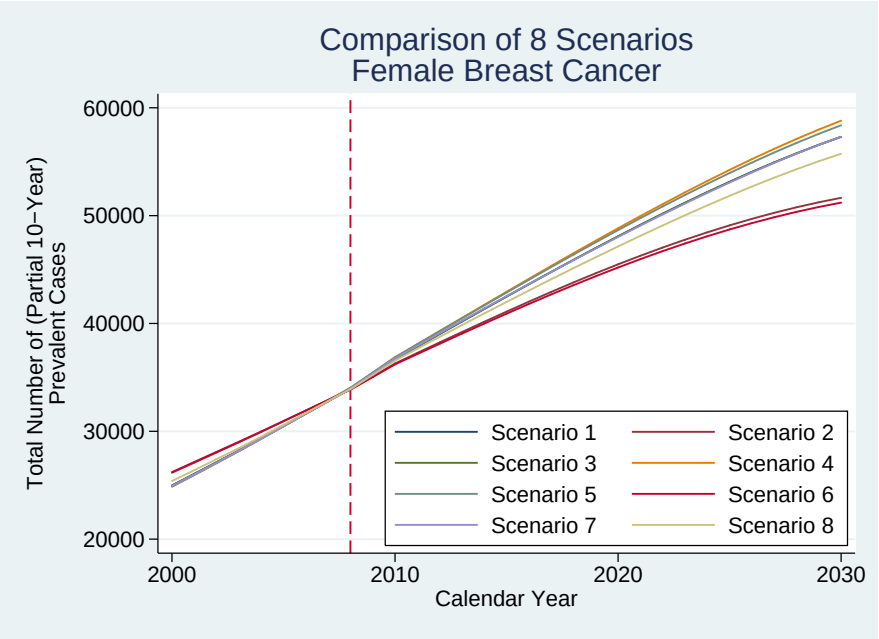

FIGURE 9.20. Comparison of the projections made under 8 Scenarios for female breast cancer data.

of 1998 compared to 1995 for the other five scenarios). The projections given in this figure were calculated using a model that does not have an interaction between age at diagnosis and year of diagnosis. It has been observed that this assumption was not reasonable for the available breast cancer data (see Figure 8.12); this can be relaxed in the projections by including an interaction term in the survival model.

As in Figure 9.9, it is possible to report the age-specific prevalence for any given calendar year from the projected prevalence estimates. Figure 9.21 compares the age-specific 10-year partial prevalence estimates for female breast cancer in Finland for the years 2008 and 2020. The projections are made under Scenario 1 of Table 9.1. Breast cancer prevalence is set to increase for the older patients if the current trends continue into the future. This is due to three reasons; firstly the population structure in 2020 is projected to have a higher proportion of elderly patients, secondly the incidence is projected to increase over calendar time, and finally, the survival is set to improve over calendar time if the current trends persist. The reason that the prevalence is not also higher for the younger patients is due to two further reasons; the projected trend for the cohort term, which is projected to decrease over calendar time and the

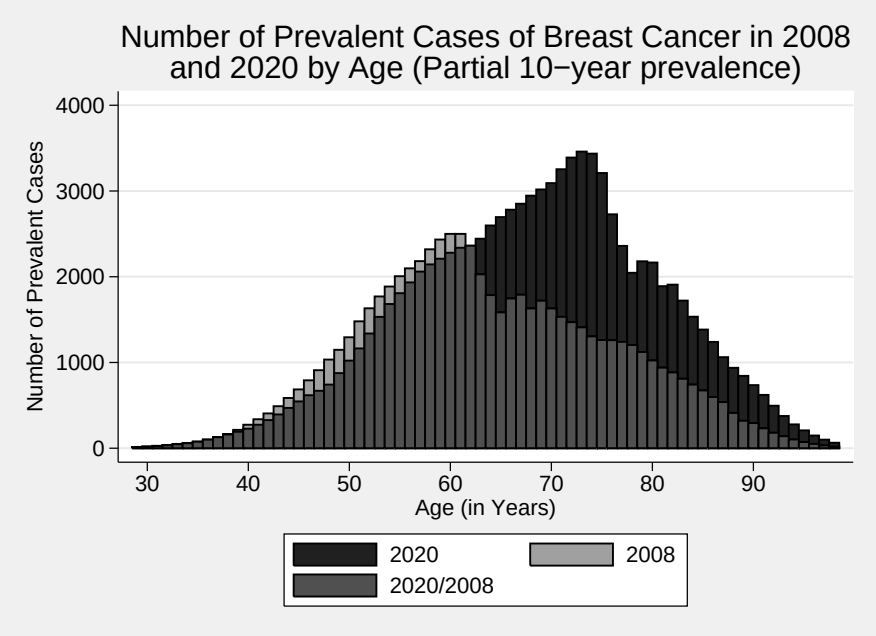

FIGURE 9.21. Comparison of the age-specific 10-year partial prevalence for female breast cancer between 2008 and 2020.

fact that the population structure in 2020 is projected to have fewer patients in the 40-60 age category due to the prospect of an ageing population [Antolin et al., 2001; Statistics Finland].

The 8 scenarios described in Table 9.1 were also applied to the male lung cancer data and the resulting projections are given in Figure 9.22. The projections are all fairly close together and give a range of possible future projections for male lung cancer under the assumption that the trends for lung cancer continue. The incidence of lung cancer was decreasing over calendar time within the observed data and the projection approaches all assume that this trend will continue into the future. Another possible scenario that could have been considered would be to assume that both the incidence and survival remain the same as the last observed interval. However, it has been shown that these estimates do not often provide good estimates of the future trends (see Figure 9.12 and 9.13 for the constant survival assumption).

9.11. Discussion

Prevalence estimates can provide a useful combined estimate of cancer burden. However, care must be taken in the definition used to define a prevalent case in order for the measure to be meaningful and of use. The partial prevalence approach appears to provide the clearest way

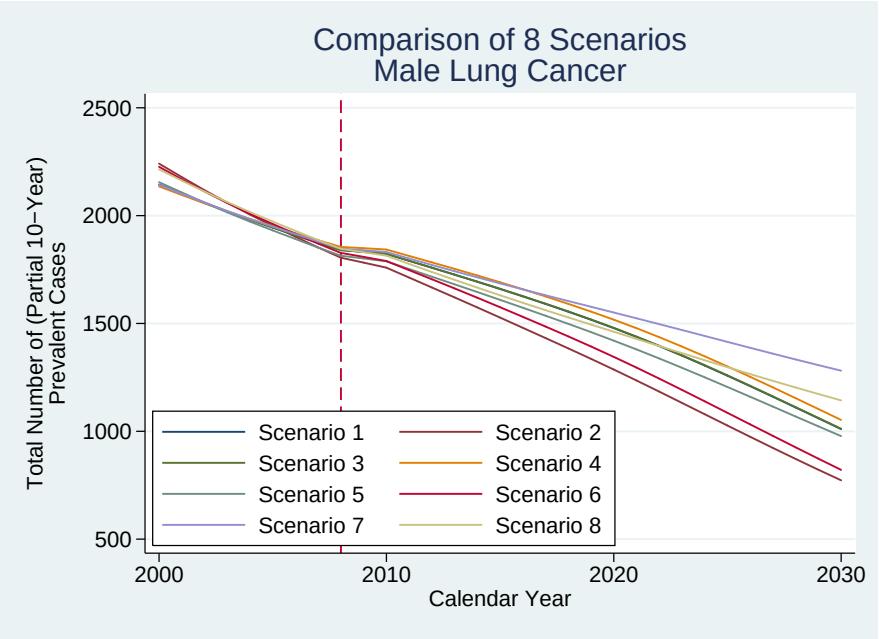

FIGURE 9.22. Comparison of the projections made under 8 Scenarios for male lung cancer data.

for the estimates to give a relevant number of patients. However, this requires a choice of how far back to go when only considering patients diagnosed less than X years ago as a prevalent case. This decision can be guided by looking at the relative survival curves for cancer patients to try to detect a plateau. Within the flexible parametric framework, it is possible to consider a direct comparison between a model that assumes statistical cure, and a model that does not. This may well be an effective way to make an improved judgement.

Combining model-based incidence and survival estimates to obtain an estimate of prevalence provides greater flexibility in the assumptions that can be made. Directly estimating prevalence via a counting method approach is possible within the range of the available data. However, projecting this estimate as opposed to projecting incidence and survival separately may fail to capture the true trends that are underlying the changes in prevalence over calendar time. Therefore, an approach that combines estimates from incidence and survival models was favoured to obtain and project the prevalence estimates.

The projected estimates of prevalence are sensitive to the modelling assumptions that are made for the incidence and survival models. This was highlighted in the projection section (Section 9.10) through the scenario-based analysis that was conducted for lung and breast

cancer. The fact that a variety of assumptions can be made and each projected estimate can be then compared provides a range of plausible future values under the different scenarios. This begins to account for the uncertainty in the assumptions that are made in order to obtain the projected estimates and is equivalent to the approach that was adopted when comparing the incidence projections in Chapter 4.

Using splines to model the effect of age leads to an improved estimation and presentation of the results from the model-based analyses. The fact that age is treated continuously means that prevalence can be reported for specific ages, and the prevalence estimates can be reported across the age-range. These estimates lead to further insight, and the data is usually available to provide yearly estimates for both age and calendar time. Comparisons between using splines and categorising age and year have been given in Chapters 4 and 8, and the disadvantages of categorising were discussed in Chapter 2. Using splines gives a more realistic representation of the effect of both age and period, whilst also providing the opportunity to provide more specific and informative projections.

CHAPTER 10

Discussion

10.1. Chapter Outline

In this chapter the thesis is concluded with a general discussion of the work, and a discussion of potential future work in the area. The limitations of the work are given consideration and a general assessment of the developed methodology is given in light of those limitations.

10.2. Introduction

There have been great improvements in the treatment and care of cancer, and there are even those who believe that curing cancer completely is achievable [Freireich, 2001]. Until that time, it is necessary for health planning authorities to prepare for the future burden of cancer on society in order to appropriately provide the resources necessary. It is also the case that many countries have the prospect of an ageing population over the coming years; this could have important implications in terms of healthcare planning for not just cancer but also other diseases associated with age. Projecting cancer rates and the survival proportion of cancer in the future is not a simple task, but it is possible to use the data that is available as a guide. Cancer registration has become standard in developed, and developing, countries and the data that is collected can be used to assess the current burden of cancer. By making simple and sensible assumptions, it is possible to project those estimates into the future.

The work carried out that comprises this thesis has been a collective effort at improving the methods for projecting the key components of cancer burden. Novel methods for projecting cancer incidence from age-period-cohort models have been proposed. This method has built upon recent work in using restricted cubic splines in order to provide smooth estimates of the incidence rate. In addition, estimates for providing up-to-date estimates of relative survival in the flexible parametric framework have also been developed and assessed.

10.3. Summary of Chapters

In the first chapter, the aims of the thesis are set out, and the key concepts are introduced. The second chapter concentrates on methods for modelling incidence; particularly age-period-cohort models. Having introduced the methodology and reviewed the literature, an approach based on restricted cubic splines is adopted and further developed before being highlighted through an example. The software in order to carry out the analyses has been developed as part of the PhD thesis [Rutherford et al., 2010]. This software has also been used by others in applied literature [Coviello et al., 2010], and recently a collaboration with IARC has started to use the sofware for an international comparison of lung cancer incidence. Making the software publicly available and the associated paper, should help the developments to be used further in practice. As part of the paper and the chapter, an approach to including interactions with key covariates is discussed and an improvement using a "reduced" set of splines is proposed.

In the third chapter, a simulation to investigate knot placement in age-period-cohort models is carried out. This work is used to highlight that the use of splines should not be discounted on the grounds of difficulty over knot choices. The work shows that in a number of settings, the number of knots selected does not overly affect the fit of the function provided that a sufficient number of knots are used. Selection criteria are also examined for the process of selecting the number of knots. The results highlight that the BIC could provide a lower bound for the number of knots to use in any given situation. Carrying out a sensitivity analysis for the number of knots used for age, period and cohort should satisfy any queries over knot selection when using splines. To carry out the simulation, fractional polynomials were used to generate the underlying shape of age and period. This choice led to some difficulties in the youngest ages, as a turning point was simulated when there was little information for the model to fully capture the shape. This was particularly an issue for the weighted knot placement. Consequently, the weighted knot placement performed poorly in the simulations. However, using a different data generation process for the simulation would have seen the weighted knot placement perform much better.

Chapter 4 introduces a new technique for making incidence projections based on the age-period-cohort model using restricted cubic splines. A paper describing the new technique has

been submitted and is currently undergoing peer-review [Rutherford et al., 2011b]. The technique is validated using a retrospective analysis and compared to established methods for incidence projection. The improvements seen from the newly proposed technique are due to two separate reasons. Firstly, improvements are seen due to the continuous representation of the age, period and cohort variables using splines; standard approaches use factor models with aggregated data. Secondly, using the restriction of the cubic splines to make the projections leads to more recent data dictating the shape of the future projections. Two sensitivity analyses are carried out to examine the robustness of the technique to the number of knots and the placement of the boundary knot.

In Chapter 5 a brief overview of the challenges for incidence projection is given. The dangers of making projections are highlighted through a number of examples using Finnish Cancer Registry data. In the chapter, discussion is given of how best to minimise the danger by using the available registry data, and incorporating external evidence. A lot of applications for incidence projections have used the same method of projection for a number of cancer sites. This approach is criticised and a more thoughtful approach to projection is advocated.

Chapter 6 is simply an introduction to the key methodology for the analysis of time-to-event data; that is, survival analysis. Flexible parametric models are introduced and the flexibility of the approach is highlighted; both in terms of capturing the baseline hazard and in terms of their application. Key concepts for the analysis of population-based cancer data; such as relative survival and period analysis, are also introduced. Finally the concept of statistical cure is given consideration due to its potential for being utilised in providing a more appropriate estimate of cancer burden.

Chapter 7 gives a comparison of the approaches for estimating relative survival. The results of the simulation highlight that adjusting for age is necessary if an estimate of net survival is required; that is, an estimate that is independent of the populations' background risk of death. The work carried out in this chapter makes recommendations of the most appropriate methods to use and has been recently published [Rutherford et al., 2011a].

Chapter 8 gives an evaluation of the modelled period analysis approach using the flexible parametric modelling framework. A comparison is also given to having an interaction between

follow-up time, and the improvement over calendar time. This approach to obtaining up-to-date, and potentially projected estimates of relative survival is coupled with the advice from the previous chapter concerning adjusting for age. Previous expositions of the modelled period analysis approach in the literature have failed to make this adjustment. On the basis that it is essential to obtain an estimate of net survival from the relative survival approaches, taking age into account is vital for modelled period analysis.

In Chapter 9, prevalence is estimated by combining the estimates obtained in the previous chapters. There is a detailed discussion of the different prevalence estimates that can be obtained. Estimates of model-based partial prevalence are shown for a number of cancer sites. There are also examples of the projected estimates of prevalence using the approaches to projection for incidence and survival covered in the earlier chapters.

10.4. Achieving the Aims of the Thesis

Throughout the chapters and developments that have been produced across the chapters of the thesis, the aims of the thesis have been met. The aims of the thesis were to estimate and project the cancer burden on society. In the specification of the aims, it was decided that prevalence would be used as a proxy for cancer burden. The other components of the cancer burden such as cost and quality of life require an accurate future projection of the number of patients that have cancer (prevalence). Using the prevalence estimates it is the possible to use other clinical registers and external information to begin to calculate the financial and social cost of cancer. In chapters 2 and 4, improvements on the estimation and projection of cancer incidence have been given through the use of restricted cubic splines. Improvements of the estimation and projection of cancer patient survival have also been developed in Chapters 7 and 8. These have then been combined in Chapter 9 to satisfy the aim of providing future estimates of cancer prevalence.

10.5. Assessment of the Proposed Methods

The proposed methods in the thesis have been validated to some extent by employing retrospective analyses for the Finnish cancer registry data. This technique which uses historical data to validate the projection approaches has been employed for each of the proposed projection methods. The data that has been used was provided by the Finnish Cancer Registry. Finland is

a relatively small country but the cancer registry data is of extremely high quality and there is also a long series of data stretching back until the 1950s. Therefore, the dataset is ideally suited for the retrospective analyses employed in this thesis. However, the proposed techniques should be employed on data from other, larger countries in order to further validate the approaches.

The methods that have been developed as part of this thesis have been applied to a wide-range of cancer sites through the examples that have been given. The main emphasis of the thesis was methodological development for projecting the cancer burden. However, the examples that have been given have also provided interesting interpretation points through both the refining of current methods and the development of new methodology. The use of the methods in the future for practical applications has been made easier by the development of user-friendly software as part of the thesis.

The proposed methodology has generally been based on the use of restricted cubic splines so that continuous representations of the quantities can be given. The age-specific estimates of survival and incidence that feed into the age-specific prevalence estimates have been presented in Chapter 9. These provide an improved understanding of the components that combine to give an estimate of prevalence, whilst also providing a greater understanding of the effect of age. Presentation of the results from statistical models is vitally important. In this thesis, graphical representation has been sought to ensure that the results of complex statistical models can be easily understood and fully interpreted.

Chapter 7 shows the results of a comparison of approaches to estimating relative survival and the main results have been recently published [Rutherford et al., 2011a]. The conclusions of the chapter and paper make recommendations for practice when using relative survival to estimate net survival and stress the importance of taking age into account. Two other recent publications [Hakulinen et al., 2011; Perme et al., 2011] have similarly given recommendations on approaches to estimating net survival when using a relative survival methodology. The three papers combined provide the necessary details on how long-established methods do not estimate the true net survival in certain situations. Taking account of age in the modelling of relative survival also has important implications for estimating projected survival estimates and the consequent projections of prevalence.

10.6. Limitations

The generalisability of the findings could be improved by applying the methods to further datasets from a variety of countries. The methods were developed using the Finnish Cancer Registry data due to its high quality and completeness. However, further application of the method to other datasets will help to refine the methods of projection.

Another limitation is the lack of focus on uncertainty of the projected estimates. There are two forms of uncertainty that affect the projections. Firstly, there is standard random variation around the estimates. It is possible to obtain this estimate of uncertainty for both the projected survival and incidence estimates using the delta method. Secondly, there is a degree of uncertainty surrounding the assumption that is made when making the projection. This uncertainty is harder to quantify, and the sensitivity analyses that have been applied for both the incidence and prevalence projections provide one method of observing this uncertainty. The resulting plots give a "fan" of projection estimates that are based on variants of the assumptions made for projecting the estimates into the future. However, further work is needed in this area and this is further discussed in the following section.

A common criticism of the use of splines is the arbitrary nature of the decision over the number and placement of the knots. Restricted cubic splines have been used extensively throughout this thesis because of a firm belief that this criticism is not important in practice provided that common sense is applied when placing the knots. Chapter 3 gives a detailed comparison for knot placement in terms of the age-period-cohort models and highlights that the fit of the function is very similar provided that a sufficient number of knots is used. Further work is necessary to investigate whether similar conclusions can be drawn when using restricted cubic splines to capture to baseline (excess) hazard when using flexible parametric models. A similar simulation study to that carried out in Chapter 3 would lend further weight to the use of the flexible parametric approach.

Data quality is a perennial issue when using population-based data. The Finnish Cancer Registry [Finnish Cancer Registry] is a long-established and internationally renowned cancer registry. The completeness of the Finnish Cancer Registry has been shown to be in excess of 99% for solid tumours [Teppo et al., 1994]. The process undertaken in Finland to ensure that patients are appropriately linked to the death registry ensures that only a small proportion of

patients are lost during follow-up. However, considerations on the data quality are important when drawing conclusions surrounding the results produced from that data.

Another consideration when using long time-series of data is changes of the definition for a disease over time. This can have important implications when trying to model trends in survival, incidence and prevalence. A related issue to this is the impact that screening can have on the three measures. Prostate cancer is a prime example of the impact that screening can have on the incidence rate [Hankey et al., 1999; Etzioni et al., 2002; Coldman et al., 2003] as well as the other two measures. Care must be taken when making projections from cancer sites where there has been a systematic change in the disease definition or if there is any reason for an excess of cases being diagnosed. Making projections from such data without extreme care is likely to lead to spurious projections. The topics highlighted in Chapter 5, particularly the "Recommendations for Practice", should be considered when making projections from population-based data.

10.7. Future Work

The newly proposed method for incidence projections requires further validation in a wide-range of settings. The results of the method applied to the Finnish cancer registry data are encouraging but further validation can only help to fine-tune the method. The sensitivity analyses carried out at the end of Chapter 4 could be further extended. However, the plots showing the various boundary knot placements appear to be a nice feature of the method. These plots can be used to begin to assess the uncertainty that is associated with making an unverifiable assumption of linearity to project into the future. It is intended that further comparisons will be undertaken in the future to compare the projected estimates from the newly proposed method to current projections for the UK [Møller et al., 2007]. Discussions have also been undertaken to use the data available to IARC to make a more widescale validation of the newly proposed method.

The work on modelled period analysis using flexible parametric models appears to provide a useful technique for obtaining up-to-date estimates of relative survival. In terms of providing an appropriate measure of net survival, the extension to include the effect of age in the modelling has provided an improvement on what is currently reported. The effect of age has been treated as a categorical variable in the first part of Chapter 8. However, there is the potential to use a continuous representation of the effect of age through the use of splines (as highlighted in

Section 8.5). Model-based age-standardisation is still possible when using splines to model the effect of age, and this is an area that could be given further consideration in the future. Single figure estimates are often required for international comparisons, and this is an area that has received a lot of recent interest in the literature [Hakulinen et al., 2011; Perme et al., 2011; Rutherford et al., 2011a].

The major consideration for future work is to attempt to provide further estimates of uncertainty around the produced estimates. There are two forms of uncertainty to consider. There is the standard random variation around the estimates, and there is the added uncertainty introduced by making an unverifiable assumption in order to make the projections.

10.8. Final Conclusions

Using appropriately split data to provide continuous estimates over age and calendar time for the key estimates of cancer burden is of benefit. This is made possible by selecting an appropriate modelling framework and employing splines to smooth estimates. The work carried out as part of this thesis illustrates and proposes methods to perform these analyses for modelling survival, incidence and prevalence. Sensible assumptions for projecting the future cancer burden can be made so that the data made available by population-based cancer registries can be utilised for projections. This is of vital importance for health authorities so that they can appropriately plan and prepare the appropriate resources in the future.

Appendix I

Appendix I contains the code for `apcfit` and `poprisktime` written for the statistical computer package Stata.

apcfit.ado

```
/*define program name*/
capture program drop apcfit
program define apcfit, rclass

/*set the syntax for the program*/
syntax [varlist(default=none)] [if] [in], Age(varname) ///
Cases(varname) POPrisktime(varname) ///
[Period(varname) AGEFitted(string) PERFitted(string) COHFitted(string) ///
REFCoh(real 0) DRExtr(string) REFPer(real 0) COHort(varname) ///
PARam(string) LEVel(int 95) DFA(int 5) DFP(int 5) DFC(int 5) NPER(int 1) ///
BKNOTSA(numlist max=2 min=2) BKNOTSP(numlist max=2 min=2) BKNOTSC(numlist max=2 min=2) ///
KNOTSA(numlist ascending) KNOTSP(numlist ascending) KNOTSC(numlist ascending) ///
RMATRIXA(name) RMATRIXP(name) RMATRIXC(name) ///
KNOTPLacement(string) ADJust LINK(string) ITERate(int 16000) replace]

marksample touse

capture drop _sp*
capture drop _drift

/*replace option for the fitted value variables. */
if "`param'"!="AC" & "`param'"!="AP" {
if "`replace'"!="" {
capture drop agefitted agefitted_lci agefitted_uci ///
perfitted perfitted_lci perfitted_uci cohfitted ///
cohfitted_lci cohfitted_uci
        if _rc==111 {
            display as error "The replace option is specified but some or all of the variables to be replaced cannot be
> found."
            exit 198
        }
}
}

if "`param'"=="AC" {
if "`replace'"!="" {
capture drop agefitted agefitted_lci agefitted_uci ///
 cohfitted cohfitted_lci cohfitted_uci
        if _rc==111 {
            display as error "The replace option is specified but some or all of the variables to be replaced cannot be
> found."
            exit 198
        }
}
}

if "`param'"!=="AP" {
if "`replace'"!="" {
capture drop agefitted agefitted_lci agefitted_uci ///
perfitted perfitted_lci perfitted_uci
        if _rc==111 {
            display as error "The replace option is specified but some or all of the variables to be replaced cannot be
> found."
            exit 198
        }
}
}

/*Checks correct variables are specified */
if "`param'"!="AC" & "`period'"=="" {
di as error "Period must be specified for all parameterisations except for AC"
exit 198
}

if "`param'"=="AC" & "`cohort'"=="" {
di as error "Cohort must be specified for the AC parameterisation"
exit 198
}
```

```
/*Checks that df and knots aren't both specified*/

if "`dfa'"!="5" & "`knotsa'"!="" {
di as error "Degrees of freedom option and knots option for Age cannot be specified simultaneously"
exit 198
}

if "`dfp'"!="5" & "`knotsp'"!="" {
di as error "Degrees of freedom option and knots option for Period cannot be specified simultaneously"
exit 198
}

if "`dfc'"!="5" & "`knotsc'"!="" {
di as error "Degrees of freedom option and knots option for Cohort cannot be specified simultaneously"
exit 198
}

/*Set local macros to make code simpler*/
local A "`age'"
local D "`cases'"
local Y "`poprisktime'"

if "`param'"!="AC" {
local P "`period'"
}

/*if statements to deal with the case when cohort is/isn't defined */
if "`cohort'"!="" {
    local C "`cohort'"
}

if "`cohort'"=="" & "`param'"!="AP" {
    tempvar C
  quietly   gen `C'=`P'-`A'
}

/*Check if rcsgen is installed or not*/
capture: which rcsgen
if _rc==111 {
        display as error "rcsgen must be installed in order to run apcfit. This can be installed by typing: ssc install rcsgen, in
>  the command window."
        exit 198
 }

if "`agefitted'"=="" {
     local agefitted "agefitted"
}

if "`perfitted'"=="" & "`param'"!="AC" {
     local perfitted "perfitted"
}

if "`cohfitted'"=="" & "`param'"!="AP" {
     local cohfitted "cohfitted"
}
/*Check if agefitted already defined*/
capture: su `agefitted'
if _rc== 0 {
        display as error "The variable being used for the fitted age values is already defined. Either drop the variable, use the
> replace option, or use the agefitted option to define a different name for the fitted values."
        exit 198
 }

capture: su `agefitted'_lci
if _rc== 0 {
        display as error "The variable being used for the fitted age values LCI is already defined. Either drop the variable, use
> the replace option, or use the agefitted option to define a different name for the fitted values."
        exit 198
 }
```

```
capture: su `agefitted'_uci
if _rc== 0 {
	display as error "The variable being used for the fitted age values UCI is already defined. Either drop the variable, use
> the replace option, or use the agefitted option to define a different name for the fitted values."
	exit 198
 }

if "`param'"!="AC" {
capture: su `perfitted'
if _rc== 0 {
	display as error "The variable being used for the fitted period values is already defined. Either drop the variable, or us
> e the perfitted option to define a different name for the fitted values."
	exit 198
 }

capture: su `perfitted'_lci
if _rc== 0 {
	display as error "The variable being used for the fitted period values LCI is already defined. Either drop the variable, o
> r use the perfitted option to define a different name for the fitted values."
	exit 198
 }

capture: su `perfitted'_uci
if _rc== 0 {
	display as error "The variable being used for the fitted period values UCI is already defined. Either drop the variable, o
> r use the perfitted option to define a different name for the fitted values."
	exit 198
 }

}

if "`param'"!="AP" {
capture: su `cohfitted'
if _rc== 0 {
	display as error "The variable being used for the fitted cohort values is already defined. Either drop the variable, or us
> e the cohfitted option to define a different name for the fitted values."
	exit 198
 }

capture: su `cohfitted'_lci
if _rc== 0 {
	display as error "The variable being used for the fitted cohort values LCI is already defined. Either drop the variable, o
> r use the cohfitted option to define a different name for the fitted values."
	exit 198
 }

capture: su `cohfitted'_uci
if _rc== 0 {
	display as error "The variable being used for the fitted cohort values UCI is already defined. Either drop the variable, o
> r use the cohfitted option to define a different name for the fitted values."
	exit 198
 }
}

/*Method for extracting the median references if refp and refc are not defined*/

if "`param'"!="AC" {
	quietly summarize `P' [aweight=`D'] if `touse',d
	quietly gen refdefp0=r(p50) if _n==1
}

if "`param'"!="AP" {
	quietly summarize `C' [aweight=`D'] if `touse',d
	quietly gen refdefc0=r(p50) if _n==1
}

if "`param'"!="AP" {
scalar define defc0=refdefc0
drop refdefc0
}
```

```
if "`param'"!="AC" {
scalar define defp0=refdefp0
drop refdefp0
}

/*Sets the median references as default if the user does not specify references*/
if "`refcoh'"=="0" & "`param'"!="AP" {
    local c0=defc0
}

if "`refper'"=="0" & "`param'"!="AC" {
    local p0=defp0
}

/*Sets the references as those defined by the user if they choose to define them*/
if "`refcoh'"!="0" {
    local c0 "`refcoh'"
}

if "`refper'"!="0" {
    local p0 "`refper'"
}

/*Sets the default drift extraction to be weighted*/
if "`drextr'"=="" {
    local drextr "weighted"
}

/*Displays an error if drextr is incorrectly specified*/
if "`drextr'"!="" & "`drextr'"!="weighted" & "`drextr'"!="holford" {
    display as error "if drextr is specified it must be specified as weighted or holford"
exit
}

/*Sets ACP as the default parameterisation*/
if "`param'"=="" {
   local param "ACP"
}

/*Displays an error if param is incorrectly specified*/
if "`param'"!="" & "`param'"!="ACP" & "`param'"!="APC" & "`param'"!="AdCP" & "`param'"!="AdPC" & "`param'"!="AP" & "`param'"!="AC" {
   display as error "if param is specified it must be specified as one of the given options"
   exit 198
}

/*Sets equal as the default parameterisation*/
if "`knotplacement'"=="" {
   local knotplacement "equal"
}

/*Displays an error if knotplacement is incorrectly specified*/
if "`knotplacement'"!="" & "`knotplacement'"!="equal" & "`knotplacement'"!="weighted" {
   display as error "if knotplacement is specified it must be specified as one of the given options"
   exit 198
}

if "`adjust'"!="" & "`param'"=="AP"  {
      display as error "Adjust should not be specified with the AP or AC parameterisations"
      exit 198
}

if "`adjust'"!="" & "`param'"=="AC"  {
      display as error "Adjust should not be specified with the AP or AC parameterisations"
      exit 198
}

/*Checks and sets the default link function*/

if "`link'"!="" & "`link'"!="log" & "`link'"!="power5" {
   display as error "if link is specified it must be specified as one of the given options"
   exit 198
```

```
}

if "`link'"=="" {
    local link "log"
}

/*Preserves the dataset whilst the MAs matrix is created in Mata*/
preserve
   quietly keep if `touse'
   gen colA0=1

      if "`rmatrixa'"!="" {
      if "`knotsa'"!="" {
            local nk : word count `knotsa'
            local dfa = `nk' - 1

                  if "`knotplacement'"=="equal" {
                        quietly rcsgen `A', gen(colA) rmatrix(`rmatrixa') knots(`knotsa') bknots(`bknotsa')
                  }
                  if "`knotplacement'"=="weighted" {
                        quietly rcsgen `A', gen(colA) rmatrix(`rmatrixa') knots(`knotsa') bknots(`bk
> notsa') fw(`D')

                  }
      }

      else {
   if "`knotplacement'"=="equal" {
            quietly rcsgen `A', gen(colA) rmatrix(`rmatrixa') df(`dfa') bknots(`bknotsa')
   }
   if "`knotplacement'"=="weighted" {
            quietly rcsgen `A', gen(colA) rmatrix(`rmatrixa') df(`dfa') bknots(`bknotsa') fw(`D')
   }
      }
      }

      else {
   if "`knotsa'"!="" {
            local nk : word count `knotsa'
            local dfa = `nk' - 1

                  if "`knotplacement'"=="equal" {
                        quietly rcsgen `A', gen(colA) orthog knots(`knotsa') bknots(`bknotsa')
                  }
                  if "`knotplacement'"=="weighted" {
                        quietly rcsgen `A', gen(colA) orthog knots(`knotsa') bknots(`bknotsa') fw(`D
> ')

                  }
      }

      else {
   if "`knotplacement'"=="equal" {
            quietly rcsgen `A', gen(colA) orthog df(`dfa') bknots(`bknotsa')
   }
   if "`knotplacement'"=="weighted" {
            quietly rcsgen `A', gen(colA) orthog df(`dfa') bknots(`bknotsa') fw(`D')
   }
      }
      }

   local aknots `r(knots)'
   matrix RmatA=r(R)
      return matrix RmatA=RmatA, copy
   mata: RmatA=st_matrix("r(R)")
   keep colA*
   mata: MAs=st_data(.,(.))
restore
preserve
   quietly keep if `touse'
   keep `A'
```

```
   mata: tA=st_data(.,("`A'"))
restore
preserve
  quietly keep if `touse'
  keep `D'
  mata: DA=st_data(.,("`D'"))
restore

if "`param'"!="AC" {

/*Preserves the dataset whilst the MPs matrix is created in Mata*/
preserve
   quietly keep if `touse'
      if "`rmatrixp'"!="" {
           if "`knotsp'"!="" {
      local nk : word count `knotsp'
      local dfp = `nk' - 1

   if "`knotplacement'"=="equal" {
            quietly rcsgen `P', gen(colP) rmatrix(`rmatrixp') knots(`knotsp') bknots(`bknotsp')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `P', gen(colP) rmatrix(`rmatrixp') knots(`knotsp') bknots(`bknotsp') fw(`D')
   }
      }

      else {
   if "`knotplacement'"=="equal" {
            quietly rcsgen `P', gen(colP) rmatrix(`rmatrixp') df(`dfp') bknots(`bknotsp')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `P', gen(colP) rmatrix(`rmatrixp') df(`dfp') bknots(`bknotsp') fw(`D')
   }
      }
      }

      else {
      if "`knotsp'"!="" {
      local nk : word count `knotsp'
      local dfp = `nk' - 1

   if "`knotplacement'"=="equal" {
            quietly rcsgen `P', gen(colP) orthog knots(`knotsp') bknots(`bknotsp')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `P', gen(colP) orthog knots(`knotsp') bknots(`bknotsp') fw(`D')
   }
      }

      else {
   if "`knotplacement'"=="equal" {
            quietly rcsgen `P', gen(colP) orthog df(`dfp') bknots(`bknotsp')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `P', gen(colP) orthog df(`dfp') bknots(`bknotsp') fw(`D')
   }
      }
      }

      if "`rmatrixp'"!="" {
   local pknots `r(knots)'
   matrix RmatP=`rmatrixp'
      return matrix RmatP=RmatP, copy
      }
      else {
      local pknots `r(knots)'
   matrix RmatP=r(R)
      return matrix RmatP=RmatP, copy
      }

   mata: RmatP=st_matrix("r(R)")
```

```
  keep colP*
  mata: MPs=st_data(.,(.))
restore
preserve
  quietly keep if `touse'
  keep `P'
  mata: tP=st_data(.,("`P'"))
restore
preserve
  quietly keep if `touse'
  keep `D'
  mata: DP=st_data(.,("`D'"))
restore
}


if "`param'"!="AP" {

/*Preserves the dataset whilst the MCs matrix is created in Mata*/
preserve
   quietly keep if `touse'

      if "`rmatrixc'"!="" {
      if "`knotsc'"!="" {
      local nk : word count `knotsc'
      local dfc = `nk' - 1
      if "`knotplacement'"=="equal" {
            quietly rcsgen `C', gen(colC) rmatrix(`rmatrixc') knots(`knotsc') bknots(`bknotsc')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `C', gen(colC) rmatrix(`rmatrixc') knots(`knotsc') bknots(`bknotsc') fw(`D')
   }
      }
      else {
   if "`knotplacement'"=="equal" {
            quietly rcsgen `C', gen(colC) rmatrix(`rmatrixc') df(`dfc') bknots(`bknotsc')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `C', gen(colC) rmatrix(`rmatrixc') df(`dfc') bknots(`bknotsc') fw(`D')
   }
      }
      }


      else {
if "`knotsc'"!="" {
      local nk : word count `knotsc'
      local dfc = `nk' - 1
      if "`knotplacement'"=="equal" {
            quietly rcsgen `C', gen(colC) orthog knots(`knotsc') bknots(`bknotsc')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `C', gen(colC) orthog knots(`knotsc') bknots(`bknotsc') fw(`D')
   }
      }
      else {
   if "`knotplacement'"=="equal" {
            quietly rcsgen `C', gen(colC) orthog df(`dfc') bknots(`bknotsc')
   }
   if "`knotplacement'"=="weighted" {
           quietly rcsgen `C', gen(colC) orthog df(`dfc') bknots(`bknotsc') fw(`D')
   }
      }
      }

      if "`rmatrixc'"!="" {
   local cknots `r(knots)'
   matrix RmatC=`rmatrixc'
      return matrix RmatC=RmatC, copy
      }
      else {
```

```
      local cknots `r(knots)'
   matrix RmatC=r(R)
      return matrix RmatC=RmatC, copy
      }

   mata: RmatC=st_matrix("r(R)")
   keep colC*
   mata: MCs=st_data(.,(.))
restore
preserve
  quietly keep if `touse'
  keep `C'
  mata: tC=st_data(.,("`C'"))
restore
preserve
  quietly keep if `touse'
  keep `D'
  mata: DC=st_data(.,("`D'"))
restore
}

if "`param'"=="AP" {
tempvar p0col
tempvar dfAcol
tempvar dfPcol
quietly gen `p0col'=`p0'
quietly gen `dfAcol'=`dfa'
quietly gen `dfPcol'=`dfp'

mata: dfa=st_data(1,"`dfAcol'")
mata: dfp=st_data(1,"`dfPcol'")
mata: p0=st_data(1,"`p0col'")

      preserve
        quietly keep if `touse'
        quietly rcsgen `p0col', gen(colp0) rmatrix(RmatP) knots(`pknots')
        keep colp0*
        mata: RP=st_data(.,(.))
        mata: RP=RP[1,.]
      restore

mata: P0=progP0matrix(RP,MPs)
mata: MPs0=MPs-P0

/*Generates the correct number of empty Stata variables for the Age coeffs*/
local x=1
while `x'<=`dfa'+1 {
  quietly gen _spA`x'=.
  local x=`x'+1
}

  local y=1
  while `y'<=`dfp' {
    quietly gen _spP`y'=.
    local y=`y'+1
}

order _spP*
mata: usedrows=rows(MPs0)

mata: for (j=1;j<=dfp; j++)  st_store(.,(j),"`touse'",MPs0[.,j])

/*Orders the variables to allow easier storage of the columns from Mata*/
order _spA*
/*Stores the results in Stata from Mata*/
mata: for (i=1; i<=dfa+1; i++) st_store(.,(i),"`touse'",MAs[.,i])

/*Generates variables showing the unique groups for the variables*/
tempvar grA
tempvar grP
quietly egen `grA'=group(`A') if `touse'
```

```
quietly egen `grP'=group(`P') if `touse'

/*Generates a variable that is 1s and 0s that tag the first element of the unique groups*/
tempvar tagA
tempvar tagP
quietly egen `tagA'=tag(`A') if `touse'
quietly egen `tagP'=tag(`P') if `touse'

/*Generates an id variable*/
tempvar _id
quietly egen `_id'=seq() if `touse'

/*Generates a variable that shows the value and position of the unique values of the variables*/
tempvar Apos
tempvar Ppos
quietly gen `Apos'=`tagA'*`grA' if `touse'
quietly gen `Ppos'=`tagP'*`grP' if `touse'

/*Replaces the 0s as missing*/
quietly replace `Apos'=. if `Apos'==0
quietly replace `Ppos'=. if `Ppos'==0

/*Generates unique values of the variables and their position in the matrices*/
tempvar Aposact
quietly gen `Aposact'=`_id' if `Apos'!=.
quietly ta `A' if `touse', matrow(uniqueA)
mata: uniqueA=st_matrix("uniqueA")
mata: rowsUA=rows(uniqueA)
mata: Apos=st_data(.,"`Aposact'")
mata: Apos=editvalue(Apos,.,0)
mata: Apos=select(Apos, Apos[.,1]:>0)

/*Generates unique values of the variables and their position in the matrices*/
tempvar Pposact
quietly gen `Pposact'=`_id' if `Ppos'!=.
quietly ta `P' if `touse', matrow(uniqueP)

mata: uniqueP=st_matrix("uniqueP")
mata: rowsUP=rows(uniqueP)
mata: Ppos=st_data(.,"`Pposact'")
mata: Ppos=editvalue(Ppos,.,0)
mata: Ppos=select(Ppos, Ppos[.,1]:>0)

rename _spA1 _spA1_intct

if "`link'"!="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* if `touse', lnoffset(`Y') f(p) nocons
}
else {
glm `D' _sp* if `touse', lnoffset(`Y') f(p) nocons itetate(`iterate')
}
}
if "`link'"=="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* if `touse', f(p) nocons link(power5 `Y')
}
else {
glm `D' _sp* if `touse', f(p) nocons link(power5 `Y') iterate(`iterate')

}
}

rename _spA1_intct _spA1

  mata: coeffs=st_matrix("e(b)")
  mata: betaA=coeffs'[(1..dfa+1),.]
  mata: betaP=coeffs'[(dfa+2..(dfa+dfp+1)),.]

  mata: varcov=st_matrix("e(V)")
  mata: varcovA=varcov[(1..dfa+1),(1..dfa+1)]
```

```
  mata: varcovP=varcov[(dfa+2..(dfa+dfp+1)),(dfa+2..(dfa+dfp+1))]

  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=MPs0[Ppos,.]

/*Creates matrices containing the fitted values*/
mata: answerA=exp(cutdownA*betaA)
mata: answerP=exp(cutdownP*betaP)

/*Creates matrices containing the fitted cov/vars*/
mata: varianceA=cutdownA*varcovA*transposeonly(cutdownA)
mata: varianceP=cutdownP*varcovP*transposeonly(cutdownP)

/*Creates matrices containing just the vars*/
mata: varianceA=diagonal(varianceA)
mata: varianceP=diagonal(varianceP)

/*Generates the appropriate z value for the normal distribution for the CIs*/
local alpha2 = (100-`level')/200
local zalpha2 = -invnorm(`alpha2')

/*Calculates the 95% (or user defined level) CIs*/
mata: UCIA=exp(ln(answerA)+`zalpha2':*sqrt(varianceA))
mata: LCIA=exp(ln(answerA)-`zalpha2':*sqrt(varianceA))

mata: UCIP=exp(ln(answerP)+`zalpha2':*sqrt(varianceP))
mata: LCIP=exp(ln(answerP)-`zalpha2':*sqrt(varianceP))

/*Generates the variables for the answers to store into*/
quietly gen `agefitted'=.
quietly gen `agefitted'_lci=.
quietly gen `agefitted'_uci=.
quietly gen `perfitted'=.
quietly gen `perfitted'_lci=.
quietly gen `perfitted'_uci=.

/*Stores the answers into Stata variables*/

quietly replace `Aposact'=0 if `Aposact'==.
mata: st_store(.,"`agefitted'","`Aposact'",answerA)
mata: st_store(.,"`agefitted'_lci","`Aposact'",LCIA)
mata: st_store(.,"`agefitted'_uci","`Aposact'",UCIA)

quietly replace `Pposact'=0 if `Pposact'==.
mata: st_store(.,"`perfitted'","`Pposact'",answerP)
mata: st_store(.,"`perfitted'_lci","`Pposact'",LCIP)
mata: st_store(.,"`perfitted'_uci","`Pposact'",UCIP)

quietly replace `agefitted'=`nper'*`agefitted'
quietly replace `agefitted'_lci=`nper'*`agefitted'_lci
quietly replace `agefitted'_uci=`nper'*`agefitted'_uci

scalar define la=`dfa'+1
local la=la

foreach r of numlist `dfp'/1 {
  move _spP`r' `perfitted'_uci
}

foreach t of numlist `la'/1 {
  move _spA`t' `perfitted'_uci
}

return local refper `p0'

return local knotsAge `aknots'
return local boundknotsAge `abknots'

return local knotsPer `pknots'
return local boundknotsPer `pbknots'
```

```
rename _spA1 _spA1_intct
}

if "`param'"=="AC" {
tempvar c0col
tempvar dfAcol
tempvar dfCcol
quietly gen `c0col'=`c0'
quietly gen `dfAcol'=`dfa'
quietly gen `dfCcol'=`dfc'

mata: dfa=st_data(1,"`dfAcol'")
mata: dfc=st_data(1,"`dfCcol'")
mata: c0=st_data(1,"`c0col'")

      preserve
        quietly keep if `touse'
        quietly rcsgen `c0col', gen(colc0) rmatrix(RmatC) knots(`cknots')
        keep colc0*
        mata: RC=st_data(.,(.))
        mata: RC=RC[1,.]
      restore

mata: C0=progC0matrix(RC,MCs)
mata: MCs0=MCs-C0

/*Generates the correct number of empty Stata variables for the Age coeffs*/
local x=1
while `x'<=`dfa'+1 {
  quietly gen _spA`x'=.
  local x=`x'+1
}

  local z=1
  while `z'<=`dfc' {
    quietly gen _spC`z'=.
    local z=`z'+1
}

order _spC*

mata: usedrows=rows(MCs0)

mata: for (j=1;j<=dfc; j++)  st_store(.,(j),"`touse'",MCs0[.,j])

/*Orders the variables to allow easier storage of the columns from Mata*/
order _spA*
/*Stores the results in Stata from Mata*/
mata: for (i=1; i<=dfa+1; i++) st_store(.,(i),"`touse'",MAs[.,i])

/*Generates variables showing the unique groups for the variables*/
tempvar grA
tempvar grC
quietly egen `grA'=group(`A') if `touse'
quietly egen `grC'=group(`C') if `touse'

/*Generates a variable that is 1s and 0s that tag the first element of the unique groups*/
tempvar tagA
tempvar tagC
quietly egen `tagA'=tag(`A') if `touse'
quietly egen `tagC'=tag(`C') if `touse'

/*Generates an id variable*/
tempvar _id
quietly egen `_id'=seq() if `touse'

/*Generates a variable that shows the value and position of the unique values of the variables*/
tempvar Apos
tempvar Cpos
quietly gen `Apos'=`tagA'*`grA' if `touse'
quietly gen `Cpos'=`tagC'*`grC' if `touse'
```

```
/*Replaces the 0s as missing*/
quietly replace `Apos'=. if `Apos'==0
quietly replace `Cpos'=. if `Cpos'==0

/*Generates unique values of the variables and their position in the matrices*/
tempvar Aposact
quietly gen `Aposact'=`_id' if `Apos'!=.
quietly ta `A' if `touse', matrow(uniqueA)
mata: uniqueA=st_matrix("uniqueA")
mata: rowsUA=rows(uniqueA)
mata: Apos=st_data(.,"`Aposact'")
mata: Apos=editvalue(Apos,.,0)
mata: Apos=select(Apos, Apos[.,1]:>0)

/*Generates unique values of the variables and their position in the matrices*/
tempvar Cposact
quietly gen `Cposact'=`_id' if `Cpos'!=.
quietly ta `C' if `touse', matrow(uniqueC)

mata: uniqueC=st_matrix("uniqueC")
mata: rowsUC=rows(uniqueC)
mata: Cpos=st_data(.,"`Cposact'")
mata: Cpos=editvalue(Cpos,.,0)
mata: Cpos=select(Cpos, Cpos[.,1]:>0)

rename _spA1 _spA1_intct

if "`link'"!="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* if `touse', lnoffset(`Y') f(p) nocons
}
else {
glm `D' _sp* if `touse', lnoffset(`Y') f(p) nocons iterate(`iterate')
}
}
if "`link'"=="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* if `touse', f(p) nocons link(power5 `Y')
}
else {
glm `D' _sp* if `touse', f(p) nocons link(power5 `Y') iterate(`iterate')
}
}

rename _spA1_intct _spA1

  mata: coeffs=st_matrix("e(b)")
  mata: betaA=coeffs'[(1..dfa+1),.]
  mata: betaC=coeffs'[(dfa+2..(dfa+dfc+1)),.]

  mata: varcov=st_matrix("e(V)")
  mata: varcovA=varcov[(1..dfa+1),(1..dfa+1)]
  mata: varcovC=varcov[(dfa+2..(dfa+dfc+1)),(dfa+2..(dfa+dfc+1))]

  mata: cutdownA=MAs[Apos,.]
  mata: cutdownC=MCs0[Cpos,.]

/*Creates matrices containing the fitted values*/
mata: answerA=exp(cutdownA*betaA)
mata: answerC=exp(cutdownC*betaC)

/*Creates matrices containing the fitted cov/vars*/
mata: varianceA=cutdownA*varcovA*transposeonly(cutdownA)
mata: varianceC=cutdownC*varcovC*transposeonly(cutdownC)
/*Creates matrices containing just the vars*/
mata: varianceA=diagonal(varianceA)
mata: varianceC=diagonal(varianceC)

/*Generates the appropriate z value for the normal distribution for the CIs*/
local alpha2 = (100-`level')/200
```

```
local zalpha2 = -invnorm(`alpha2')

/*Calculates the 95% (or user defined level) CIs*/
mata: UCIA=exp(ln(answerA)+`zalpha2':*sqrt(varianceA))
mata: LCIA=exp(ln(answerA)-`zalpha2':*sqrt(varianceA))

mata: UCIC=exp(ln(answerC)+`zalpha2':*sqrt(varianceC))
mata: LCIC=exp(ln(answerC)-`zalpha2':*sqrt(varianceC))

/*Generates the variables for the answers to store into*/
quietly gen `agefitted'=.
quietly gen `agefitted'_lci=.
quietly gen `agefitted'_uci=.
quietly gen `cohfitted'=.
quietly gen `cohfitted'_lci=.
quietly gen `cohfitted'_uci=.

/*Stores the answers into Stata variables*/

quietly replace `Aposact'=0 if `Aposact'==.
mata: st_store(.,"`agefitted'","`Aposact'",answerA)
mata: st_store(.,"`agefitted'_lci","`Aposact'",LCIA)
mata: st_store(.,"`agefitted'_uci","`Aposact'",UCIA)

quietly replace `Cposact'=0 if `Cposact'==.
mata: st_store(.,"`cohfitted'","`Cposact'",answerC)
mata: st_store(.,"`cohfitted'_lci","`Cposact'",LCIC)
mata: st_store(.,"`cohfitted'_uci","`Cposact'",UCIC)

quietly replace `agefitted'=`nper'*`agefitted'
quietly replace `agefitted'_lci=`nper'*`agefitted'_lci
quietly replace `agefitted'_uci=`nper'*`agefitted'_uci

foreach r of numlist `dfc'/1 {
  move _spC`r' `cohfitted'_uci
}

scalar define la=`dfa'+1
local la=la

foreach t of numlist `la'/1 {
  move _spA`t' `cohfitted'_uci
}

return local refcoh `c0'
return local knotsAge `aknots'
return local boundknotsAge `abknots'
return local knotsCoh `cknots'
return local boundknotsCoh `cbknots'

rename _spA1 _spA1_intct

}

if "`param'"!="AP" & "`param'"!="AC" {

/*Generates variables that are later dropped*/
tempvar c0col
tempvar p0col
quietly gen `c0col'=`c0'
quietly gen `p0col'=`p0'

tempvar dfAcol
tempvar dfPcol
tempvar dfCcol
quietly gen `dfAcol'=`dfa'
quietly gen `dfPcol'=`dfp'
quietly gen `dfCcol'=`dfc'

/*Sets the values of c0 and p0 as Mata 1x1 matrices*/
mata: c0=st_data(1,"`c0col'")
```

```
mata: p0=st_data(1,"`p0col'")
mata: dfa=st_data(1,"`dfAcol'")
mata: dfp=st_data(1,"`dfPcol'")
mata: dfc=st_data(1,"`dfCcol'")

/*Generates the row required for the reference cohort*/
      preserve
        quietly keep if `touse'
        quietly rcsgen `c0col', gen(colc0) rmatrix(RmatC) knots(`cknots')
        keep colc0*
        mata: RC=st_data(.,(.))
        mata: RC=RC[1,.]
      restore

/*Generates the row required for the reference period*/

      preserve
        quietly keep if `touse'
        quietly rcsgen `p0col', gen(colp0) rmatrix(RmatP) knots(`pknots')
        keep colp0*
        mata: RP=st_data(.,(.))
        mata: RP=RP[1,.]
      restore

/*Generates the matrices with the added rows for the references*/
mata: MPsplusrow=(RP\MPs)
mata: MCsplusrow=(RC\MCs)
mata: tPplusrow=(p0\tP)
mata: tCplusrow=(c0\tC)

/*Performs the detrending with a weighted drift extraction*/
if "`drextr'"=="weighted" | "`drextr'"=="" {
mata: detrendMPsplusrow=detrendMfinalweighted(MPsplusrow,tPplusrow,DP)
mata: detrendMCsplusrow=detrendMfinalweighted(MCsplusrow,tCplusrow,DC)
}

/*Performs the detrending with a Holford drift extraction (w=col(1))*/
if "`drextr'"=="holford" {
mata: detrendMPsplusrow=detrendMfinalholford(MPsplusrow,tPplusrow,DP)
mata: detrendMCsplusrow=detrendMfinalholford(MCsplusrow,tCplusrow,DC)
}

/*Performs the manipulations to calculate the adjusted detrended matrices*/
mata: rowC0=detrendMCsplusrow[(1),.]
mata: C0=progC0matrix(rowC0,detrendMCsplusrow)
mata: nrowC0=rows(C0)
mata: C0=C0[2..nrowC0,.]
mata: nrowdMCspr=rows(detrendMCsplusrow)
mata: detrendMC0=detrendMCsplusrow[2..nrowdMCspr,.]-C0
mata detrendMCn=detrendMCsplusrow[2..nrowdMCspr,.]
mata: ColC=tC
mata: ColCminC0=ColC:-c0
mata: detrendMCfinal=(ColCminC0,detrendMC0)
mata: detrendMCfinalNA=(ColC,detrendMCn)


/*Performs the manipulations to calculate the adjusted detrended matrices*/
mata: rowP0=detrendMPsplusrow[(1),.]
mata: P0=progP0matrix(rowP0,detrendMPsplusrow)
mata: nrowP0=rows(P0)
mata: P0=P0[2..nrowP0,.]
mata: nrowdMPspr=rows(detrendMPsplusrow)
mata: detrendMP0=detrendMPsplusrow[2..nrowdMPspr,.]-P0
mata detrendMPn=detrendMPsplusrow[2..nrowdMPspr,.]
mata: ColP=tP
mata: ColPminP0=ColP:-p0
mata: detrendMPfinal=(ColPminP0,detrendMP0)
mata: detrendMPfinalNA=(ColP,detrendMPn)
/*Generates the correct number of empty Stata variables for the Age coeffs*/
local x=1
while `x'<=`dfa'+1 {
```

```
  quietly gen _spA`x'=.
  local x=`x'+1
}

/*Generates the correct number of empty Stata variables for the Period coeffs*/
if "`param'"=="ACP" | "`param'"=="AdCP" | "`param'"=="AdPC" {
  local y=1
  while `y'<=`dfp'-1 {
    quietly gen _spP`y'=.
    local y=`y'+1
}
}

if "`param'"=="APC" {
  local y=1
  while `y'<=`dfp' {
    quietly gen _spP`y'=.
    local y=`y'+1
}
}

/*Generates the correct number of empty Stata variables for the Cohort coeffs*/
if "`param'"=="ACP"  {
  local z=1
  while `z'<=`dfc' {
    quietly gen _spC`z'=.
    local z=`z'+1
}
}

if "`param'"=="APC" | "`param'"=="AdPC" | "`param'"=="AdCP" {
  local z=1
  while `z'<=`dfc'-1 {
    quietly gen _spC`z'=.
    local z=`z'+1
}
}

if "`adjust'"!="" {
/*Orders the variables to allow easier storage of the columns from Mata*/
order _spC*

mata: usedrows=rows(detrendMCfinal)

/*Stores the results in Stata from Mata*/


if "`param'"=="ACP"  {
  mata: for (j=1;j<=dfc; j++)  st_store(.,(j),"`touse'",detrendMCfinal[.,j])
}

if "`param'"=="APC" | "`param'"=="AdPC" | "`param'"=="AdCP" {
  mata: for (j=1; j<=dfc-1; j++)  st_store(.,(j),"`touse'",detrendMC0[.,j])
}

/*Orders the variables to allow easier storage of the columns from Mata*/
order _spP*
/*Stores the results in Stata from Mata*/
if "`param'"=="ACP" | "`param'"=="AdCP" | "`param'"=="AdPC" {
  mata: for (k=1; k<=dfp-1; k++) st_store(.,(k),"`touse'",detrendMP0[.,k])
}

if "`param'"=="APC"  {
  mata: for (k=1; k<=dfp; k++) st_store(.,(k),"`touse'",detrendMPfinal[.,k])
}
}

if "`adjust'"=="" {

/*Orders the variables to allow easier storage of the columns from Mata*/
order _spC*
```

```
mata: usedrows=rows(detrendMCfinal)

/*Stores the results in Stata from Mata*/
if "`param'"=="ACP"  {
  mata: for (j=1;j<=dfc; j++)  st_store(.,(j),"`touse'",detrendMCfinal[.,j])
}

if  "`param'"=="AdCP" {
  mata: for (j=1; j<=dfc-1; j++)  st_store(.,(j),"`touse'",detrendMC0[.,j])
}


if "`param'"=="APC" | "`param'"=="AdPC" {
  mata: for (j=1; j<=dfc-1; j++)  st_store(.,(j),"`touse'",detrendMCn[.,j])
}

/*Orders the variables to allow easier storage of the columns from Mata*/
order _spP*
/*Stores the results in Stata from Mata*/
if "`param'"=="ACP" | "`param'"=="AdCP" | "`param'"=="AdPC" {
  mata: for (k=1; k<=dfp-1; k++) st_store(.,(k),"`touse'",detrendMPn[.,k])
}

if "`param'"=="AdPC" {
  mata: for (k=1; k<=dfp-1; k++) st_store(.,(k),"`touse'",detrendMP0[.,k])
}

if "`param'"=="APC"  {
   mata: for (k=1; k<=dfp; k++) st_store(.,(k),"`touse'",detrendMPfinal[.,k])
}
}

/*Orders the variables to allow easier storage of the columns from Mata*/
order _spA*
/*Stores the results in Stata from Mata*/
mata: for (i=1; i<=dfa+1; i++) st_store(.,(i),"`touse'",MAs[.,i])

/*Generates variables showing the unique groups for the variables*/
tempvar grA
tempvar grP
tempvar grC
quietly egen `grA'=group(`A') if `touse'
quietly egen `grP'=group(`P') if `touse'
quietly egen `grC'=group(`C') if `touse'

/*Generates a variable that is 1s and 0s that tag the first element of the unique groups*/
tempvar tagA
tempvar tagP
tempvar tagC
quietly egen `tagA'=tag(`A') if `touse'
quietly egen `tagP'=tag(`P') if `touse'
quietly egen `tagC'=tag(`C') if `touse'

/*Generates an id variable*/
tempvar _id
quietly egen `_id'=seq() if `touse'

/*Generates a variable that shows the value and position of the unique values of the variables*/
tempvar Apos
tempvar Ppos
tempvar Cpos
quietly gen `Apos'=`tagA'*`grA' if `touse'
quietly gen `Ppos'=`tagP'*`grP' if `touse'
quietly gen `Cpos'=`tagC'*`grC' if `touse'

/*Replaces the 0s as missing*/
quietly replace `Apos'=. if `Apos'==0
quietly replace `Cpos'=. if `Cpos'==0
quietly replace `Ppos'=. if `Ppos'==0
```

```
/*Generates unique values of the variables and their position in the matrices*/

tempvar Aposact
quietly gen `Aposact'=`_id' if `Apos'!=.
quietly ta `A' if `touse', matrow(uniqueA)
mata: uniqueA=st_matrix("uniqueA")
mata: rowsUA=rows(uniqueA)
mata: Apos=st_data(.,"`Aposact'")
mata: Apos=editvalue(Apos,.,0)
mata: Apos=select(Apos, Apos[.,1]:>0)

/*Generates unique values of the variables and their position in the matrices*/

tempvar Pposact
quietly gen `Pposact'=`_id' if `Ppos'!=.
quietly ta `P' if `touse', matrow(uniqueP)

mata: uniqueP=st_matrix("uniqueP")
mata: rowsUP=rows(uniqueP)
mata: Ppos=st_data(.,"`Pposact'")
mata: Ppos=editvalue(Ppos,.,0)
mata: Ppos=select(Ppos, Ppos[.,1]:>0)

/*Generates unique values of the variables and their position in the matrices*/

tempvar Cposact
quietly gen `Cposact'=`_id' if `Cpos'!=.
quietly ta `C' if `touse', matrow(uniqueC)

mata: uniqueC=st_matrix("uniqueC")
mata: rowsUC=rows(uniqueC)
mata: Cpos=st_data(.,"`Cposact'")
mata: Cpos=editvalue(Cpos,.,0)
mata: Cpos=select(Cpos, Cpos[.,1]:>0)


/*Carries out the glm with the appropriate offset, and poisson dist for the detrended splines*/

if "`param'"=="APC" {
  rename _spP1 _spP1_ldrft
}

if "`param'"=="ACP" {
  rename _spC1 _spC1_ldrft
}

rename _spA1 _spA1_intct

if "`param'"!="AdPC" & "`param'"!="AdCP" {
if "`link'"!="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* if `touse', lnoffset(`Y') f(p) nocons
}
else {
glm `D' _sp* if `touse', lnoffset(`Y') f(p) nocons iterate(`iterate')
}
}
if "`link'"=="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* if `touse', f(p) nocons link(power5 `Y')
}
else {
glm `D' _sp* if `touse', f(p) nocons link(power5 `Y') iterate(`iterate')
}
}
}

if "`param'"=="AdPC" {
gen _drift=`P'-`p0'
if "`link'"!="power5" {
if "`iterate'"=="16000" {
```

```
glm `D' _sp* _drift if `touse', lnoffset(`Y') f(p) nocons
}
else {
glm `D' _sp* _drift if `touse', lnoffset(`Y') f(p) nocons iterate(`iterate')
}
}
if "`link'"=="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* _drift if `touse', f(p) nocons link(power5 `Y')
}
else {
glm `D' _sp* _drift if `touse', f(p) nocons link(power5 `Y') iterate(`iterate')
}
}
}

if "`param'"=="AdCP" {
gen _drift=`C'-`c0'
if "`link'"!="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* _drift if `touse', lnoffset(`Y') f(p) nocons
}
else {
glm `D' _sp* _drift if `touse', lnoffset(`Y') f(p) nocons iterate(`iterate')
}
}
if "`link'"=="power5" {
if "`iterate'"=="16000" {
glm `D' _sp* _drift if `touse', f(p) nocons link(power5 `Y')
}
else {
glm `D' _sp* _drift if `touse', f(p) nocons link(power5 `Y') iterate(`iterate')
}
}
}

if "`param'"=="APC" {
  rename _spP1_ldrft _spP1
}

if "`param'"=="ACP" {
  rename _spC1_ldrft _spC1
}

rename _spA1_intct _spA1

/*Obtains the coefficients from the glm and sorts them into the appropriate variables*/
if "`param'"=="ACP" {
  mata: coeffs=st_matrix("e(b)")
  mata: betaA=coeffs'[(1..dfa+1),.]
  mata: betaP=coeffs'[(dfa+2..(dfa+dfp)),.]
  mata: betaC=coeffs'[(dfa+dfp+1)..(dfa+dfp+dfc),.]
}

if "`param'"=="AdCP" {
  mata: coeffs=st_matrix("e(b)")
  mata: betaA=coeffs'[(1..dfa+1),.]
  mata: betaP=coeffs'[(dfa+2..(dfa+dfp)),.]
  mata: betaC=coeffs'[(dfa+dfp+1)..(dfa+dfp+dfc-1),.]
}

if "`param'"=="APC" {
  mata: coeffs=st_matrix("e(b)")
  mata: betaA=coeffs'[(1..dfa+1),.]
  mata: betaP=coeffs'[(dfa+2..(dfa+dfp+1)),.]
  mata: betaC=coeffs'[(dfa+dfp+2)..(dfa+dfp+dfc),.]
}

if "`param'"=="AdPC" {
  mata: coeffs=st_matrix("e(b)")
  mata: betaA=coeffs'[(1..dfa+1),.]
```

```
  mata: betaP=coeffs'[(dfa+2..(dfa+dfp)),.]
  mata: betaC=coeffs'[(dfa+dfp+1)..(dfa+dfp+dfc-1),.]
}

/*Obtains the vars/covs from the glm and sorts them into the appropriate variables*/

if "`param'"=="ACP" {
  mata: varcov=st_matrix("e(V)")
  mata: varcovA=varcov[(1..dfa+1),(1..dfa+1)]
  mata: varcovP=varcov[(dfa+2..(dfa+dfp)),(dfa+2..(dfa+dfp))]
  mata: varcovC=varcov[((dfa+dfp+1)..(dfa+dfp+dfc)),((dfa+dfp+1)..(dfa+dfp+dfc))]
}

if "`param'"=="AdCP" {
  mata: varcov=st_matrix("e(V)")
  mata: varcovA=varcov[(1..dfa+1),(1..dfa+1)]
  mata: varcovP=varcov[(dfa+2..(dfa+dfp)),(dfa+2..(dfa+dfp))]
  mata: varcovC=varcov[((dfa+dfp+1)..(dfa+dfp+dfc-1)),((dfa+dfp+1)..(dfa+dfp+dfc-1))]
}

if "`param'"=="APC" {
  mata: varcov=st_matrix("e(V)")
  mata: varcovA=varcov[(1..dfa+1),(1..dfa+1)]
  mata: varcovP=varcov[(dfa+2..(dfa+dfp+1)),(dfa+2..(dfa+dfp+1))]
  mata: varcovC=varcov[((dfa+dfp+2)..(dfa+dfp+dfc)),((dfa+dfp+2)..(dfa+dfp+dfc))]
}

if "`param'"=="AdPC" {
  mata: varcov=st_matrix("e(V)")
  mata: varcovA=varcov[(1..dfa+1),(1..dfa+1)]
  mata: varcovP=varcov[(dfa+2..(dfa+dfp)),(dfa+2..(dfa+dfp))]
  mata: varcovC=varcov[((dfa+dfp+1)..(dfa+dfp+dfc-1)),((dfa+dfp+1)..(dfa+dfp+dfc-1))]
}

/*Creates matrices corresponding to the unique values of the variables*/
if "`adjust'"!="" {
if "`param'"=="ACP" {
  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=detrendMP0[Ppos,.]
  mata: cutdownC=detrendMCfinal[Cpos,.]
}

if "`param'"=="AdCP" | "`param'"=="AdPC" {
  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=detrendMP0[Ppos,.]
  mata: cutdownC=detrendMC0[Cpos,.]
}

if "`param'"=="APC" {
  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=detrendMPfinal[Ppos,.]
  mata: cutdownC=detrendMC0[Cpos,.]
}
}

if "`adjust'"=="" {
if "`param'"=="ACP" {
  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=detrendMPn[Ppos,.]
  mata: cutdownC=detrendMCfinal[Cpos,.]
}

if "`param'"=="AdCP"  {
  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=detrendMPn[Ppos,.]
  mata: cutdownC=detrendMC0[Cpos,.]
}

if "`param'"=="AdPC" {
  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=detrendMP0[Ppos,.]
```

```
  mata: cutdownC=detrendMCn[Cpos,.]
}

if "`param'"=="APC" {
  mata: cutdownA=MAs[Apos,.]
  mata: cutdownP=detrendMPfinal[Ppos,.]
  mata: cutdownC=detrendMCn[Cpos,.]
}
}

/*Creates matrices containing the fitted values*/
mata: answerA=exp(cutdownA*betaA)
mata: answerP=exp(cutdownP*betaP)
mata: answerC=exp(cutdownC*betaC)

/*Creates matrices containing the fitted cov/vars*/
mata: varianceA=cutdownA*varcovA*transposeonly(cutdownA)
mata: varianceP=cutdownP*varcovP*transposeonly(cutdownP)
mata: varianceC=cutdownC*varcovC*transposeonly(cutdownC)

/*Creates matrices containing just the vars*/
mata: varianceA=diagonal(varianceA)
mata: varianceP=diagonal(varianceP)
mata: varianceC=diagonal(varianceC)

/*Generates the appropriate z value for the normal distribution for the CIs*/
local alpha2 = (100-`level')/200
local zalpha2 = -invnorm(`alpha2')

/*Calculates the 95% (or user defined level) CIs*/
mata: UCIA=exp(ln(answerA)+`zalpha2':*sqrt(varianceA))
mata: LCIA=exp(ln(answerA)-`zalpha2':*sqrt(varianceA))

mata: UCIP=exp(ln(answerP)+`zalpha2':*sqrt(varianceP))
mata: LCIP=exp(ln(answerP)-`zalpha2':*sqrt(varianceP))

mata: UCIC=exp(ln(answerC)+`zalpha2':*sqrt(varianceC))
mata: LCIC=exp(ln(answerC)-`zalpha2':*sqrt(varianceC))

/*Generates the variables for the answers to store into*/
quietly gen `agefitted'=.
quietly gen `agefitted'_lci=.
quietly gen `agefitted'_uci=.
quietly gen `perfitted'=.
quietly gen `perfitted'_lci=.
quietly gen `perfitted'_uci=.
quietly gen `cohfitted'=.
quietly gen `cohfitted'_lci=.
quietly gen `cohfitted'_uci=.

/*Stores the answers into Stata variables*/

quietly replace `Aposact'=0 if `Aposact'==.
mata: st_store(.,"`agefitted'","`Aposact'",answerA)
mata: st_store(.,"`agefitted'_lci","`Aposact'",LCIA)
mata: st_store(.,"`agefitted'_uci","`Aposact'",UCIA)

quietly replace `Pposact'=0 if `Pposact'==.
mata: st_store(.,"`perfitted'","`Pposact'",answerP)
mata: st_store(.,"`perfitted'_lci","`Pposact'",LCIP)
mata: st_store(.,"`perfitted'_uci","`Pposact'",UCIP)

quietly replace `Cposact'=0 if `Cposact'==.
mata: st_store(.,"`cohfitted'","`Cposact'",answerC)
mata: st_store(.,"`cohfitted'_lci","`Cposact'",LCIC)
mata: st_store(.,"`cohfitted'_uci","`Cposact'",UCIC)

quietly replace `agefitted'=`nper'*`agefitted'
quietly replace `agefitted'_lci=`nper'*`agefitted'_lci
quietly replace `agefitted'_uci=`nper'*`agefitted'_uci
```

```
/*Drops unwanted variables*/
/*Puts the spline variables to the end of the dataset*/

if "`param'"=="ACP" {
  foreach r of numlist `dfc'/1 {
  move _spC`r' `cohfitted'_uci
}

scalar define lp=`dfp'-1
scalar define la=`dfa'+1
local lp=lp
local la=la

foreach s of numlist `lp'/1 {
   move _spP`s' `cohfitted'_uci
}

foreach t of numlist `la'/1 {
   move _spA`t' `cohfitted'_uci
}
}

if "`param'"=="APC" {
  scalar define lc=`dfc'-1
  scalar define la=`dfa'+1
  local lc=lc
  local la=la

foreach s of numlist `lc'/1 {
  move _spC`s' `cohfitted'_uci
}

foreach r of numlist `dfp'/1 {
  move _spP`r' `cohfitted'_uci
}

foreach t of numlist `la'/1 {
  move _spA`t' `cohfitted'_uci
}
}

if "`param'"=="APC" {
  rename _spP1 _spP1_ldrft
}

if "`param'"=="ACP" {
  rename _spC1 _spC1_ldrft
}

rename _spA1 _spA1_intct

return local refcoh `c0'
return local refper `p0'
return local knotsAge `aknots'
return local boundknotsAge `abknots'
return local knotsPer `pknots'
return local boundknotsPer `pbknots'
return local knotsCoh `cknots'
return local boundknotsCoh `cbknots'

}

end

/*MATA PROGRAMS*/

/*Mata program used to generate the pivot vector used later in detrend*/
mata:
real matrix progpivotvector(real matrix userv,userPM) {
v=userv
PM=userPM
```

```
rank=rank(PM,tol=-1e-03)
/* if (rank==.) { */
/* exit(_error(3200,"Missing values encountered. Restrict the data using if/in to produce an answer from apcfit.")) */
/* } */

for (j=1; j<=rank; j++) {
                v[,j]=1
      }
return(v)
}

end

/*Mata program used to detrend the matrix using the weighted method*/
mata:
real matrix detrendMfinalweighted(real matrix userM, usert, userD)
{
/*Sets the matrices to be used during the program*/
D=userD
t=usert
M=userM
numeric matrix w, X, PM, Pp, H, R1, tau, p1, v, detrendM, pivot, pivotedPM
real scalar rank

/*Generates a constant term of the right length and the X matrix*/
cons=J(rows(M),1,1)
X=(cons,t)

/*Generates the weights and sets the first row of the weights to zero*/
w=D
w=(0\D)


/*Performs  the detrending process*/
Pp=svsolve(cross(X:*sqrt(w),X:*sqrt(w)),transposeonly(X:*w))*M
PM=X*Pp
PM=M-PM

/*Generates a blank pivot vector and performs QR decompostion*/
p1=.
H=tau=R1=.
hqrdp(PM,H,tau,R1,p1)

/*Sets the pivot vector equal to that given by the decomposition*/
pivot=p1

/*Generates a vector of zeros with lenth cols(M)*/
v=J(1,cols(M),0)

/*Pivots the detrended matrix according to the defined pivot from QR*/
pivotedPM=PM[.,pivot]
/*Generates a vector that contains 1s according to the rank of PM*/
v=progpivotvector(v,PM)

/*Selects a matrix of full rank using v*/
detrendM=select(pivotedPM,v)
/*Returns the full rank detrended matrix*/
return(detrendM)
}
end

/*Mata program used to detrend the matrix using the weighted method*/
/*Same comments as above*/
mata:

real matrix detrendMfinalholford(real matrix userM, usert, userD)
{

D=userD
t=usert
M=userM
```

```
numeric matrix w, X, PM, Pp, H, tau, R1, p1, v, detrendM, pivot, pivotedPM
real scalar rank

w=J(rows(M),1,1)
X=(w,t)
w[1,1]=0

Pp=svsolve(cross(X:*sqrt(w),X:*sqrt(w)),transposeonly(X:*w))*M
PM=X*Pp
PM=M-PM

p1=.
H=tau=R1=.
hqrdp(PM,H,tau,R1,p1)

pivot=p1

v=J(1,cols(M),0)

pivotedPM=PM[.,pivot]
v=progpivotvector(v,PM)

detrendM=select(pivotedPM,v)
return(detrendM)

}

end


/*Mata program used to make C0, a matrix with a repeated row of specified length*/
mata:

real matrix progC0matrix(real matrix userc0,userdetrendMC)
 {
c0=userc0
detrendMC=userdetrendMC
Jmat=J(rows(detrendMC),cols(detrendMC),0)
for (j=1; j<=cols(detrendMC); j++) {
            for (i=1; i<=rows(detrendMC); i++) {
               Jmat[i,.]=c0
            }
      }
return(Jmat)
}
end

*/Mata program used to make P0, a matrix with a repeated row of specified length*/
mata:
real matrix progP0matrix(real matrix userp0,userdetrendMP)
 {
p0=userp0
detrendMP=userdetrendMP
Jmat=J(rows(detrendMP),cols(detrendMP),0)
for (j=1; j<=cols(detrendMP); j++) {
            for (i=1; i<=rows(detrendMP); i++) {
               Jmat[i,.]=p0
            }
      }
return(Jmat)
}

end
```

poprisktime.ado

```
capture program drop poprisktime
program define poprisktime, rclass
syntax [varlist(default=none)] using/, Age(varname) PERiod(varname) COHort(varname) Cases(varname) AGEMAX(int) AGEMIN(int) PERMIN(int) PERMAX(int)
[ POP(string) POPRISKtime(string) COVariates(varlist) MISSingreplace]

capture drop _UpperCohort
capture: assert `age'==int(`age')
if _rc==9 {
di as error "The variable for the Age values must contain only integer values of Age. It may be necessary to round the variable."
exit 198
}

capture: assert `period'==int(`period')
if _rc==9 {
di as error "The variable for the Period values must contain only integer values of Period. It may be necessary to round the variable."
exit 198
}

capture: assert `cohort'==int(`cohort')
if _rc==9 {
di as error "The variable for the Cohort values must contain only integer values of Cohort. It may be necessary to round the variable."
exit 198
}

capture: assert `cases'==int(`cases')
if _rc==9 {
di in green "Warning: The variable for the Cases values contains non-integer values."
}

if "`poprisktime'"=="" {
local poprisktime "Y"
}

if "`saving'"=="" {
local saving "APCdata"
}

if "`pop'"=="" {
local pop pop
}

capture: su `poprisktime'

if _rc!=111 {
di as error "The variable name that is specified as the population risk-time is already defined."
exit 198
}

confirm file `"`using'.dta"'

tempname cohortcheck
quietly gen `cohortcheck'=`period'-`age'
quietly gen _UpperCohort=`cohortcheck'-`cohort'
quietly drop if _UpperCohort<0

preserve
use "`using'", clear
capture: assert `age'==int(`age')
if _rc==9 {
di as error "The variable for the Age values must contain only integer values of Age. It may be necessary to round the variable."
exit 198
}
capture: assert `period'==int(`period')
if _rc==9 {
di as error "The variable for the Period values must contain only integer values of Period. It may be necessary to round the variable."
exit 198
}
```

```
quietly su
local length=r(N)
quietly append using "`using'"
quietly gen _UpperCohort=0
quietly replace _UpperCohort=1 if _n>`length'
tempfile temppop
quietly sa "`temppop'", replace
restore

* capture: merge `covariates' _UpperCohort `age' `period' using "`temppop'", sort

capture: merge 1:1 `covariates' _UpperCohort `age' `period' using "`temppop'"
if _rc==459 {
quietly drop _UpperCohort
 di as error "variables _UpperCohort A P do not uniquely identify observations in the master data. This is likely to be due to the fact that the dataset is split by a covariate and the covariate option was not specified."
exit 459
}

confirm variable `pop'

quietly sort  `covariates'  _UpperCohort  `age'  `period'
quietly gen `poprisktime'=0

quietly drop if `age'<`agemin'-1
quietly drop if `age'>`agemax'+1
quietly drop if `period'<`permin'
quietly drop if P>`permax'+1

if "`missingreplace'"!="" {
quietly replace `cases'=0 if `cases'==.
}

quietly egen max=max(`period')
quietly egen min=min(`period')
quietly gen diff=max-min+2
quietly egen maxdiff=max(diff)
quietly gen maxdiffminus1=maxdiff-1
quietly replace `poprisktime'=(1/3)*`pop'[_n]+(1/6)*`pop'[_n+maxdiff] if _UpperCohort==1
quietly replace `poprisktime'=(1/6)*`pop'[_n-maxdiffminus1]+(1/3)*`pop'[_n+1] if  _UpperCohort==0
quietly replace `poprisktime'=. if `period'==max & _UpperCohort==1
quietly drop maxdiff diff min max maxdiffminus1

quietly replace `age'=`age'+0.333 if _UpperCohort==0
quietly replace `period'=`period'+0.667 if _UpperCohort==0
quietly replace `age'=`age'+0.667 if _UpperCohort==1
quietly replace `period'=`period'+0.333 if _UpperCohort==1
quietly replace `cohort'=`period'-`age'
quietly drop _UpperCohort _merge `pop'

quietly drop if `age'<`agemin'
quietly drop if `age'>`agemax'
quietly drop if `period'<`permin'
quietly drop if `period'>`permax'

 if `poprisktime'==. {
     di in green "Warning: variable `poprisktime' contain missing values. The ranges of the age and period values in the population dataset are not appropriate to carry out the formulae for the specified age and period range given in the command. See the help file for a more detailed explanation."
    }
if `cases'==. {
di in yellow "Warning: The number of cases variable now contains missing values. It is likely that these missing values should be replaced with 0s because there are no cases of the disease in that particular age-period combination. However, if missing data was present before the merge then this may be inappropriate. The missingreplace option can be used to specify a different action."
}

end
```

Appendix II

Appendix II contains a copy of the Stata Journal article that was published in 2010 [Rutherford et al., 2010] to describe the commands written to carry out age-period-cohort modelling in Stata.

Due to third party copyright restrictions the following published article has been removed from the electronic version of this thesis:

M. J. Rutherford, P. C. Lambert, and J. R. Thompson. Age-period-cohort modeling. Stata Journal, 10(4):606-627, 2010.

The unabridged version can be consulted, on request, at the University of Leicester's David Wilson Library.

Appendix III

Appendix III contains a draft of the paper describing the new technique for making incidence projections that is currently undergoing peer-review [Rutherford et al., 2011b].

Due to third party copyright restrictions the following published article has been removed from the electronic version of this thesis:

M. J. Rutherford, J. R. Thompson, and P. C. Lambert. Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines. Statistics in Medicine (submitted), 2011b.

The unabridged version can be consulted, on request, at the University of Leicester's David Wilson Library.

Appendix IV

Appendix IV contains the paper describing the simulation to compare methods of estimating relative survival [Rutherford et al., 2011a] that was published in 2011.

Due to third party copyright restrictions the following published article has been removed from the electronic version of this thesis:

M. J. Rutherford, P. W. Dickman, and P. C. Lambert. Comparison of methods for calculating relative survival in population-based studies, June 2011a.

The unabridged version can be consulted, on request, at the University of Leicester's David Wilson Library.

Bibliography

O. Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726, July 1978.

M. Agha, B. DiMonte, M. Greenberg, C. Greenberg, R. Barr, and J. R. McLaughlin. Incidence trends and projections for childhood cancer in Ontario. *Int J Cancer*, 118(11):2809–2815, Jun 2006.

H. Akaike. *Information theory and an extension of the maximum likelihood principle*, volume 1, pages 267–281. Akademiai Kiado, 1973.

K. S. Albain, W. E. Barlow, S. Shak, G. N. Hortobagyi, R. B. Livingston, I.-T. Yeh, P. Ravdin, R. Bugarini, F. L. Baehner, N. E. Davidson, G. W. Sledge, E. P. Winer, C. Hudis, J. N. Ingle, E. A. Perez, K. I. Pritchard, L. Shepherd, J. R. Gralow, C. Yoshizawa, D. C. Allred, C. K. Osborne, and D. F. Hayes. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *The Lancet Oncology*, 11(1):55 – 65, 2010.

T. Andersson, P. Dickman, S. Eloranta, and P. Lambert. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC Medical Research Methodology*, 11(1):96, 2011.

P. Antolin, H. Oxley, and W. Suyker. How will ageing affect Finland? OECD Economics Department Working Papers 295, OECD Publishing, 2001.

M. Arbyn, A. O. Raifu, P. Autier, and J. Ferlay. Burden of cervical cancer in Europe: estimates for 2004. *Ann Oncol*, 18(10):1708–1715, Oct 2007.

A. Baker and I. Bray. Bayesian projections: what are the effects of excluding data from younger age groups? *Am J Epidemiol*, 162(8):798–805, Oct 2005.

S. A. Bashir and J. Estève. Projecting cancer incidence and mortality using Bayesian age-period-cohort models. *Journal of Epidemiology and Biostatistics*, 6(3):287–296, 2001.

A.-K. Beelte, R. Pritzkuleit, and A. Katalinic. Lung cancer incidence and mortality: current trends and projections based on data from Schleswig-Holstein. *Dtsch Med Wochenschr*, 133 (28-29):1487–1492, Jul 2008.

C. B. Begg and D. Schrag. Attribution of deaths following cancer treatment. *Journal of the National Cancer Institute*, 94(14):1044–1045, July 2002.

R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

V. Beral and R. Peto. UK cancer survival statistics. *BMJ*, 341:–, 2010.

R. Bergström, H.-O. Adami, M. Möhner, W. Zatonski, H. Storm, A. Ekbom, S. Tretli, L. Teppo, O. Akre, and T. Hakulinen. Increase in testicular cancer incidence in six European countries: a birth cohort phenomenon. *Journal of the National Cancer Institute*, 88(11):727–733, 1996.

C. Berzuini and D. Clayton. Bayesian analysis of survival on multiple time scales. *Statist. Med.*, 13(8):823–838, 1994.

J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41, Feb. 1995.

C. Bosetti, C. Bianchi, E. Negri, and C. L. Vecchia. Estimates of the incidence and prevalence of renal cell carcinoma in Italy in 2002 and projections for the years 2007 and 2012. *Tumori*, 95(2):142–145, 2009.

F. Bray and B. Møller. Predicting the future burden of cancer. *Nat Rev Cancer*, 6(1):63–74, Jan 2006.

I. Bray, P. Brennan, and P. Boffetta. Recent trends and future projections of lymphoid neoplasms–a Bayesian age-period-cohort analysis. *Cancer Causes Control*, 12(9):813–820, Nov 2001.

H. Brenner. Up-to-date survival curves of children with cancer by period analysis. *British Journal of Cancer*, 88(11):1693–1697, 2003.

H. Brenner and O. Gefeller. An alternative approach to monitoring cancer patient survival. *Cancer*, 78(9):2004–2010, 1996.

H. Brenner and O. Gefeller. Deriving more up-todate estimates of long-term patient survival. *J Clin Epidemiol*, 50(2):211–216, Feb. 1997.

H. Brenner and T. Hakulinen. Up-to-date long-term survival curves of patients with cancer by period analysis. *Journal of Clinical Oncology*, 20(3):826–832, 2002.

H. Brenner and T. Hakulinen. On crude and age-adjusted relative survival rates. *J Clin Epidemiol*, 56(12):1185–1191, Dec 2003.

H. Brenner and T. Hakulinen. Are patients diagnosed with breast cancer before age 50 years ever cured? *Journal of Clinical Oncology*, 22(3):432–438, Feb. 2004.

H. Brenner and T. Hakulinen. Up-to-date estimates of cancer patient survival even with common latency in cancer registration. *Cancer Epidemiol Biomarkers Prev*, 15(9):1727–1732, Sep 2006a.

H. Brenner and T. Hakulinen. Up-to-date and precise estimates of cancer patient survival: Model-based period analysis. *American Journal of Epidemiology*, 164(7):689–696, 2006b.

H. Brenner and T. Hakulinen. Period *versus* cohort modeling of up-to-date cancer survival. *International Journal of Cancer*, 122(4):898–904, 2008.

H. Brenner and T. Hakulinen. Up-to-date cancer survival: Period analysis and beyond. *International Journal of Cancer*, 124(6):1384–1390, 2009.

H. Brenner, B. Söderman, and T. Hakulinen. Use of period analysis for providing more up-to-date estimates of long-term survival rates: empirical evaluation among 370,000 cancer patients in Finland. *Int J Epidemiol*, 31(2):456–462, Apr 2002.

H. Brenner, V. Arndt, O. Gefeller, and T. Hakulinen. An alternative approach to age adjustment of cancer survival rates. *European Journal of Cancer*, 40(15):2317 – 2322, 2004a.

H. Brenner, O. Gefeller, and T. Hakulinen. Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer*, 40(3):326 – 335, 2004b.

H. Brenner, A. Gondos, and V. Arndt. Recent major progress in long-term cancer patient survival disclosed by modeled period analysis. *Journal of Clinical Oncology*, 25(22):3274–3280, Aug. 2007.

H. Brenner, A. Gondos, and D. Pulte. Expected long-term survival of patients diagnosed with multiple myeloma in 2006-2010. *Haematologica*, page haematol.13782, 2009a.

H. Brenner, A. Gondos, and D. Pulte. Survival expectations of patients diagnosed with Hodgkin's lymphoma in 2006-2010. *Oncologist*, 14(8):806–813, Aug 2009b.

H. Brenner, A. M. Bouvier, R. Foschi, M. Hackl, I. K. Larsen, V. Lemmens, L. Mangone, S. Francisci, and the EUROCARE Working Group. Progress in colorectal cancer survival in Europe, from the late 1980s to the early 21st century: The EUROCARE study. *Int J Cancer*, May 2011.

N. E. Breslow and N. E. Day. Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases*, 28(5-6):289 – 303, 1975.

K. P. Burnham and D. R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research*, 33:261–304, 2004.

R. Capocaccia and R. De Angelis. Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine*, 16(4):425–440, 1997.

R. Capocaccia, R. D. Angelis, L. Frova, G. Gatta, M. Sant, A. Micheli, F. Berrino, E. Conti, L. Gaf, L. Roncucci, and A. Verdecchia. Estimation and projections of colorectal cancer trends in Italy. *Int J Epidemiol*, 26(5):924–932, Oct 1997.

B. Carstensen. Age-period-cohort models for the Lexis diagram. *Statistics in Medicine*, 26: 3018–3045, 2007.

B. Carstensen, M. Plummer, E. Laara, and M. Hills. The Epi package. Technical report, R, May 2008.

A. Cayuela, S. Rodríguez-Domínguez, M. Ruiz-Borrego, and M. Gili. Age-period-cohort analysis of breast cancer mortality rates in Andalucia (Spain). *Annals of Oncology*, 15(4):686–688, Apr. 2004.

J. M. Chan, R. M. Jou, and P. R. Carroll. The relative impact and future burden of prostate cancer in the United States. *J Urol*, 172(5 Pt 2):S13–6; discussion S17, Nov 2004.

M. Chauvenet, C. Lepage, V. Jooste, V. Cottet, J. Faivre, and A.-M. Bouvier. Prevalence of patients with colorectal cancer requiring follow-up or active treatment. *European Journal of Cancer*, 45(8):1460 – 1465, 2009.

D. Clayton and E. Schifflers. Models for temporal variations in cancer rates. I: Age-period and age-cohort models. *Statistics in Medicine*, 6:449–467, 1987a.

D. Clayton and E. Schifflers. Models for temporal variations in cancer rates. II: Age-period-cohort models. *Statistics in Medicine*, 6:469–481, 1987b.

M. S. Clements, B. K. Armstrong, and S. H. Moolgavkar. Lung cancer rate predictions using generalized additive models. *Biostatistics*, 6(4):576–589, Oct 2005.

M. S. Clements, T. Hakulinen, and S. H. Moolgavkar. Re: "Bayesian projections: what are the effects of excluding data from younger age groups?". *Am J Epidemiol*, 164(3):292–3; author reply 293–4, Aug 2006.

R. Cleries, J. M. Martínez, J. Valls, L. Pareja, L. Esteban, R. Gispert, V. Moreno, J. Ribes, and J. M. Borràs. Life expectancy and age-period-cohort effects: analysis and projections of mortality in Spain between 1977 and 2016. *Public Health*, 123(2):156–162, Feb 2009.

A. J. Coldman, M. L. McBride, and T. Braun. Calculating the prevalence of cancer. *Statist. Med.*, 11(12):1579–1589, 1992.

A. J. Coldman, N. Phillips, and T. A. Pickles. Trends in prostate cancer incidence and mortality: an analysis of mortality change by screening intensity. *Canadian Medical Association Journal*, 168(1):31–35, Jan. 2003.

M. Coleman, D. Forman, H. Bryant, J. Butler, B. Rachet, C. Maringe, U. Nur, E. Tracey, M. Coory, J. Hatcher, C. McGahan, D. Turner, L. Marrett, M. Gjerstorff, T. Johannesen, J. Adolfsson, M. Lambe, G. Lawrence, D. Meechan, E. Morris, R. Middleton, J. Steward, and M. Richards. Cancer survival in Australia, Canada, Denmark, Norway, Sweden, and the UK, 1995-2007 (the International Cancer Benchmarking Partnership): An analysis of population-based cancer registry data, Jan. 2011.

D. Collett. *Modelling Survival Data in Medical Research, Second Edition.* Chapman and Hall/CRC, Mar. 2003.

M. Colonna, G. Hedelin, J. Esteve, P. Grosclaude, G. Launoy, A. Buemi, P. Arveux, B. Tretarre, G. Chaplain, J. M. Lesec'h, N. Raverdy, P. M. Carli, F. Menegoz, and J. Faivre. National cancer prevalence estimation in France. *International Journal of Cancer*, 87(2):301–304, 2000.

V. H. Coupland, C. Okello, E. A. Davies, F. Bray, and H. Møller. The future burden of cancer in London compared with England. *J Public Health (Oxf)*, 32(1):83–89, Mar 2010.

E. Coviello, G. Caputi, D. Martinelli, C. A. Germinario, and R. Prato. Mortality trends for primary liver cancer in Puglia, Italy. *European Journal of Cancer Prevention*, 19(6):417–423 10.1097/CEJ.0b013e32833ad36e, 2010.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, Jan. 1972.

R. De Angelis, R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine*, 18(4):441–454, 1999.

A. De Carli and C. La Vecchia. Age, period and cohort models: review of knowledge and implementation in GLIM. *Rev Stat App.*, 20::397–409, 1987.

E. de Vries, L. V. van de Poll-Franse, W. J. Louwman, F. R. de Gruijl, and J. W. W. Coebergh. Predictions of skin cancer incidence in the Netherlands up to 2015. *Br J Dermatol*, 152(3): 481–488, Mar 2005.

P. W. Dickman and H.-O. Adami. Interpreting trends in cancer patient survival. *J Intern Med*, 260(2):103–117, Aug 2006.

P. W. Dickman, A. Sloggett, M. Hills, and T. Hakulinen. Regression models for relative survival. *Statistics in Medicine*, 23:51–64, 2004.

K. Doksum and J.-Y. Koo. On spline estimators and prediction intervals in nonparametric regression. *Computational Statistics & Data Analysis*, 35(1):67 – 82, 2000.

S. Durrelman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8:551–561, 1989.

T. Dyba and T. Hakulinen. Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Statistics in Medicine*, 19:1741–1752, 2000.

T. Dyba and T. Hakulinen. Do cancer predictions work? *European Journal of Cancer*, 44(3): 448 – 453, 2008.

F. Ederer and H. Heise. Instructions to IBM 650 programmers in processing survival computations. Methodogical note No. 10, End Results Evaluation Section, National Cancer Institute, Bethseda MD, 1959.

F. Ederer, L. Axtell, and C. SJ. The relative survival rate: A statistical methodology. *National Cancer Institute Monograph.*, 6:101–121, 1961.

S. Eloranta, P. Lambert, N. Cavalli-Bjorkman, T.-L. Andersson, B. Glimelius, and P. Dickman. Does socioeconomic status influence the prospect of cure from colon cancer - a population-based study in Sweden 1965-2000, Nov. 2010.

A. Engeland, T. Haldorsen, S. Tretli, T. Hakulinen, L. G. Hörte, T. Luostarinen, K. Magnus, G. Schou, H. Sigvaldason, and H. H. Storm. Prediction of cancer incidence in the Nordic countries up to the years 2000 and 2010. A collaborative study of the five Nordic Cancer Registries. *APMIS Suppl*, 38:1–124, 1993.

G. Engholm, J. Ferlay, N. Christensen, F. Bray, M. L. Gjerstorff, Å. Klint, J. E. Ktlum, E. Ólafsdóttir, E. Pukkala, and H. H. Storm. NORDCAN: Cancer Incidence, Mortality and Prevalence in the Nordic Countries, Version 3.4. Association of Nordic Cancer Registries. Danish Cancer Society., 2009.

J. Estève, E. Benhamou, M. Croasdale, and L. Raymond. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine*, 9(5):529–538, 1990.

R. Etzioni, D. F. Penson, J. M. Legler, D. di Tommaso, R. Boer, P. H. Gann, and E. J. Feuer. Overdiagnosis due to prostate-specific antigen screening: Lessons from U.S. prostate cancer incidence trends. *Journal of the National Cancer Institute*, 94(13):981–990, July 2002.

A. R. Feldman, L. Kessler, M. H. Myers, and M. D. Naughton. The prevalence of cancer. *New England Journal of Medicine*, 315(22):1394–1397, Nov. 1986.

J. Ferlay, P. Autier, M. Boniol, M. Heanue, M. Colombet, and P. Boyle. Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol*, 18(3):581–592, Mar 2007.

J. Ferlay, H. Shin, F. Bray, D. Forman, C. Mathers, and D. Parkin. Globocan 2008 v1.2, cancer incidence and mortality worldwide: IARC CancerBase No. 10 [internet]., 2010.

S. E. Fienberg and W. M. Mason. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology*, 1979.

Finnish Cancer Registry. http://www.cancer.fi/syoparekisteri/en/.

W. D. Flanders. Inaccuracies of death certificate information. *Epidemiology*, 3(1):3–5, Jan. 1992.

E. Freireich. Can we conquer cancer in the twenty-first century? *Cancer Chemotherapy and Pharmacology*, 48(0):S4–S10, July 2001.

E. L. Frome. The analysis of rates using Poisson regression models. *Biometrics*, 39(3):665–674, Sept. 1983.

G. Gatta, R. Capocaccia, F. Berrino, M. R. Ruzza, P. Contiero, and the EUROPREVAL Working Group. Colon cancer prevalence and estimation of differing care needs of colon

cancer patients. *Ann Oncol*, 15(7):1136–1142, 2004.

A. E. Gelfand and D. K. Dey. Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514, Jan. 1994.

A. Gigli, A. Mariotto, L. X. Clegg, A. Tavilla, I. Corazziari, R. Capocaccia, M. Hachey, and S. Scoppa. Estimating the variance of cancer prevalence from population-based registries. *Statistical Methods in Medical Research*, 15(3):235–253, 2006.

S. L. Glaser. Black-white differences in Hodgkin's disease incidence in the United States by age, sex, histology subtype and time. *International Journal of Epidemiology*, 20(1):68–75, Mar. 1991.

N. D. Glenn. Cohort analysts' futile quest: Statistical attempts to separate age, period and cohort effects. *American Sociological Review*, 41(5):900–904, Oct. 1976.

G. H. Golub and C. F. van Loan. *Matrix Computations, 3rd Edition*. Johns Hopkins University Press, 1996.

A. Gondos, F. Bray, T. Hakulinen, H. Brenner, and E. S. W. Group. Trends in cancer survival in 11 European populations from 1990 to 2009: a model-based analysis. *Ann Oncol*, 20(3): 564–573, Mar 2009a.

A. Gondos, B. Holleczek, M. Janssen-Heijnen, D. H. Brewster, F. Bray, S. Rosso, T. Hakulinen, and H. Brenner. Model-based projections for deriving up-to-date cancer survival estimates: An international evaluation. *International Journal of Cancer*, 125(11):2666–2672, 2009b.

P. Gordon, F. Artaud, A. Aouba, F. Laurent, V. Meininger, and A. Elbaz. Changing mortality for motor neuron disease in France (1968-2007): an age-period-cohort analysis. *European Journal of Epidemiology*, pages 1–9, 2011. 10.1007/s10654-011-9595-0.

M. E. Gore, C. L. Griffin, B. Hancock, P. M. Patel, L. Pyle, M. Aitchison, N. James, R. T. Oliver, J. Mardiak, T. Hussain, R. Sylvester, M. K. Parmar, P. Royston, and P. F. Mulders. Interferon alfa-2a versus combination therapy with interferon alfa-2a, interleukin-2, and fluorouracil in patients with untreated metastatic renal cell carcinoma (mrc re04/eortc gu 30012): an open-label randomised trial. *The Lancet*, 375(9715):641 – 648, 2010.

U. S. Govindarajulu, D. Spiegelman, S. W. Thurston, B. Ganguli, and E. A. Eisen. Comparing smoothing techniques in Cox models for exposure-response relationships. *Statistics in Medicine*, 26:3725–2752, 2007.

U. S. Govindarajulu, E. J. Malloy, B. Ganguli, D. Spiegelman, and E. A. Eisen. The comparison of alternative smoothing methods for fitting non-linear exposure-response relationships with Cox models in a simulation study. *Int J Biostat*, 5(1):Article2, 2009.

B. G. Greenberg, J. J. Wright, and C. G. Sheps. A technique for analyzing some factors affecting the incidence of syphilis. *Journal of the American Statistical Association*, 45(251):373–399, Sept. 1950.

T. Hakulinen. On long-term relative survival rates. *Journal of Chronic Diseases*, 30(7):431 – 443, 1977.

T. Hakulinen. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 38(4):933–942, Dec., 1982.

T. Hakulinen and L. Tenkanen. Regression analysis of relative survival rates. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):pp. 309–317, 1987.

T. Hakulinen, L. Teppo, and E. Sax'en. Do the predictions for cancer incidence come true? Experience from Finland. *Cancer*, 57(12):2454–2458, 1986.

T. Hakulinen, K. Sepp, and P. C. Lambert. Choosing the relative survival method for cancer survival estimation. *European Journal of Cancer*, In Press, Corrected Proof:–, 2011.

B. F. Hankey, E. J. Feuer, L. X. Clegg, R. B. Hayes, J. M. Legler, P. C. Prorok, L. A. Ries, R. M. Merrill, and R. S. Kaplan. Cancer surveillance series: Interpreting trends in prostate cancer - Part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. *Journal of the National Cancer Institute*, 91(12):1017–1024, June 1999.

R. Harding, L. Selman, G. Agupio, N. Dinat, J. Downing, L. Gwyther, T. Mashao, K. Mmoledi, L. M. Sebuyira, B. Ikin, and I. J. Higginson. The prevalence and burden of symptoms amongst cancer patients attending palliative care in two African countries, Jan. 2011.

F. E. Harrell, K. L. Lee, and B. G. Pollock. Regression models in clinical studies: Determining relationships between predictors and response. *J Natl Cancer Inst*, 80(15):1198–1202, Oct 1988.

S. Heinävaara and T. Hakulinen. Predicting the lung cancer burden: Accounting for selection of the patients with respect to general population mortality. *Statistics in Medicine*, 25: 2967–2980, 2006.

H. Heinzl, A. Kaider, and G. Zlabinger. Assessing interactions of binary time-dependent covariates with time in Cox proportional hazards regression models using cubic spline functions. *Statist. Med.*, 15(23):2589–2601, 1996.

C. Heuer. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, 53:161–177, 1997.

D. G. Hoel, E. Ron, R. Carter, and K. Mabuchi. Influence of death certificate errors on cancer mortality trends. *Journal of the National Cancer Institute*, 85(13):1063–1068, July 1993.

T. R. Holford. The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36(2):299–305, June 1980.

T. R. Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39: 311–324, 1983.

T. R. Holford. An alternative approach to statistical age-period-cohort analysis. *J Chronic Dis*, 38(10):831–840, 1985.

T. R. Holford. Approaches to fitting age-period-cohort models with unequal intervals. *Stat Med*, 25(6):977–993, Mar 2006.

G. Hutchison and S. Shapiro. Lead time gained by diagnostic screening for breast cancer. *J Natl Cancer Inst*, 41(3):665–81–, Sept. 1968.

E. M. Ibrahim, A. A. Zeeneldin, B. B. Sadiq, and A. A. Ezzat. The present and the future of breast cancer burden in the Kingdom of Saudi Arabia. *Med Oncol*, 25(4):387–393, 2008.

D. S. James and A. D. Bull. Information on death certificates: Cause for concern? *Journal of Clinical Pathology*, 49(3):213–216, Mar. 1996.

M. L. G. Janssen-Heijnen and J.-W. W. Coebergh. The changing epidemiology of lung cancer in Europe. *Lung Cancer*, 41(3):245 – 258, 2003.

A. L. Johansson, T. M.-L. Andersson, C.-C. Hsieh, S. Cnattingius, and M. Lambe. Increased mortality in women with breast cancer detected during pregnancy and different periods postpartum. *Cancer Epidemiology Biomarkers & Prevention*, pages –, July 2011.

J. Kalseth, V. Halsteinli, T. Halvorsen, B. Kalseth, K. Anthun, M. Peltola, K. Kautiainen, U. Häkkinen, E. Medin, J. Lundgren, C. Rehnberg, B. B. Másdóttir, M. Heimisdottir, H. H. Bjarnadóttir, J. E. Køtlum, and J. Kilsmark. Costs of cancer in the Nordic countries a comparative study of health care costs and public income loss compensation payments related

to cancer in the Nordic countries in 2007. Technical report, SINTEF A19395 - Unrestricted, 2011.

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958.

N. Keiding. Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 332(1627):487–509, Sept. 1990.

N. Keiding. Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(3):371–412, 1991.

H.-J. Kim, M. P. Fay, E. J. Feuer, and D. N. Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Statist. Med.*, 19(3):335–351, 2000.

L. Knorr-Held and E. Rainer. Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics*, 2(1):109–129, Mar 2001.

D. Kuang, B. Nielsen, and J. P. Nielsen. Identification of the age-period-cohort model and the extended chain-ladder model. *Biometrika*, 95(4):979–986, 2008.

P. C. Lambert. Modeling of the cure fraction in survival studies. *Stata Journal*, 7(3):351–375, 2007.

P. C. Lambert and P. Royston. Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9(2):265–290, 2009.

P. C. Lambert, L. K. Smith, D. R. Jones, and J. L. Botha. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine*, 24(24):3871–3885, 2005.

P. C. Lambert, J. R. Thompson, C. L. Weston, and P. W. Dickman. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics*, 8(3):576–594, July 2007.

P. C. Lambert, P. W. Dickman, C. P. Nelson, and P. Royston. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statist. Med.*, 29(7-8): 885–895, 2010.

P. C. Lambert, L. Holmberg, F. Sandin, F. Bray, K. M. Linklater, A. Purushotham, D. Robinson, and H. Møller. Quantifying differences in breast cancer survival between England and

Norway, May 2011.

K.-M. Leung, R. M. Elashoff, and A. A. Afifi. Censoring issues in survival analysis. *Annu. Rev. Public. Health.*, 18(1):83–104, May 1997.

J. Maddams, D. Brewster, A. Gavin, J. Steward, J. Elliott, M. Utley, and H. Møller. Cancer prevalence in the United Kingdom: Estimates for 2008. *Br J Cancer*, 101(3):541–547, Aug 2009.

J. Maddams, D. M. Parkin, and S. C. Darby. The cancer burden in the United Kingdom in 2007 due to radiotherapy. *International Journal of Cancer*, pages n/a–n/a, 2011.

A. Mariotto, J. L. Warren, K. B. Knopf, and E. J. Feuer. The prevalence of patients with colorectal carcinoma under care in the U.S. *Cancer*, 98(6):1253–1261, 2003.

A. B. Mariotto, M. N. Wesley, K. A. Cronin, K. A. Johnson, and E. J. Feuer. Estimates of long-term survival for newly diagnosed cancer patients. *Cancer*, 106(9):2039–2050, 2006.

A. B. Mariotto, K. Robin Yabroff, Y. Shao, E. J. Feuer, and M. L. Brown. Projections of the cost of cancer care in the United States: 2010-2020. *Journal of the National Cancer Institute*, 103(2):117–128, Jan. 2011.

K. O. Mason, W. M. Mason, H. H. Winsborough, and W. K. Poole. Some methodological issues in cohort analysis of archival data. *American Sociological Review*, 38(2):242–258, Apr. 1973.

D. J. McKenzie. Disentangling age, cohort and time effects in the additive model*. *Oxford Bulletin of Economics and Statistics*, 68(4):473–495, 2006.

R. J. Q. McNally, F. E. Alexander, A. Staines, and R. A. Cartwright. A comparison of three methods of analysis for age-period-cohort models with application to incidence data on non-Hodgkins lymphoma. *International Journal of Epidemiology*, 26:32–46, 1997.

R. M. Merrill, R. Capocaccia, E. J. Feuer, and A. Mariotto. Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program. *Int. J. Epidemiol.*, 29(2):197–207, 2000.

D. Modelmog, S. Rahlenbeck, and D. Trichopoulos. Accuracy of death certificates - a population-based, complete-coverage, one-year autopsy study in East Germany. *Cancer Causes & Control*, 3(6):541–546, Nov. 1992.

N. Molinari, J.-F. Durand, and R. Sabatier. Bounded optimal knots for regression splines. *Computational Statistics & Data Analysis*, 45(2):159 – 178, 2004.

B. Møller, H. Fekjær, T. Hakulinen, H. Sigvaldason, H. H. Storm, M. Talbäck, and T. Handorsen. Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. *Statistics in Medicine*, 22:2751–2766, 2003.

H. Møller, L. Fairley, V. Coupland, C. Okello, M. Green, D. Forman, B. Møller, and F. Bray. The future burden of cancer in England: incidence and numbers of new patients in 2020. *Br J Cancer*, 96(9):1484–1488, May 2007.

H. Møller, F. Sandin, F. Bray, Å. Klint, K. M. Linklater, A. Purushotham, D. Robinson, and L. Holmberg. Breast cancer survival in England, Norway and Sweden: A population-based comparison. *International Journal of Cancer*, 9999(9999):NA, 2010.

A. S. Morrison. The effects of early treatment, lead time and length bias on the mortality experienced by cases detected by screening. *International Journal of Epidemiology*, 11(3): 261–267, Sept. 1982.

J. E. Muscat, M. G. Malkin, S. Thompson, R. E. Shore, S. D. Stellman, D. McRee, A. I. Neugut, and E. L. Wynder. Handheld cellular telephone use and risk of brain cancer. *JAMA: The Journal of the American Medical Association*, 284(23):3001–3007, Dec. 2000.

T. Nakamura. Bayesian cohort models for general cohort table analyses. *Annals of the Institute of Statistical Mathematics*, 38:353–370, 1986. 10.1007/BF02482523.

C. P. Nelson, P. C. Lambert, I. B. Squire, and D. R. Jones. Flexible parametric models for relative survival, with application in coronary heart disease. *Stat Med*, 26(30):5486–5498, Dec 2007.

W. Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, Nov. 1972.

NHS. www.nhs.uk/conditions/cancer.

R. O'Brien. The age-period-cohort conundrum as two fundamental problems. *Qual. Quant.*, pages 1–16, 2010.

A. H. Olsen, D. M. Parkin, and P. Sasieni. Cancer mortality in the United Kingdom: projections to the year 2025. *Br J Cancer*, 99(9):1549–1554, Nov 2008.

C. Osmond. Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology*, 14:124–129, 1985.

C. Osmond and M. J. Gardner. Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1(3):245–259, 1982.

D. M. Parkin. Global cancer statistics in the year 2000. *Lancet Oncol*, 2(9):533–543, Sep 2001.

D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani. Estimating the world cancer burden: Globocan 2000. *International Journal of Cancer*, 94(2):153–156, 2001.

N. Pashayan, J. Powles, C. Brown, and S. W. Duffy. Incidence trends of prostate cancer in East Anglia, before and during the era of PSA diagnostic testing. *Br J Cancer*, 95(3):398–400, July 2006.

M. P. Perme, J. Stare, and J. Estve. On estimation in relative survival. *Biometrics*, pages no–no, 2011.

N. Phillips, A. Coldman, and M. L. McBride. Estimating cancer prevalence using mixture models for cancer survival. *Statist. Med.*, 21(9):1257–1270, 2002.

P. Pisani, F. Bray, and D. M. Parkin. Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. *International Journal of Cancer*, 97(1):72–81, 2002.

M. Ploeg, K. K. H. Aben, and L. A. Kiemeney. The present and future burden of urinary bladder cancer in the world. *World J Urol*, 27(3):289–293, Jun 2009.

A. Pokhrel and T. Hakulinen. How to interpret the relative survival ratios of cancer patients. *European Journal of Cancer*, 44(17):2661 – 2667, 2008.

D. Pulte, A. Gondos, and H. Brenner. Expected long-term survival of patients diagnosed with acute myeloblastic leukemia during 2006-2010. *Ann Oncol*, 21(2):335–341, Feb 2010.

H. Putter, M. Fiocco, and R. B. Geskus. Tutorial in biostatistics: competing risks and multi-state models. *Statist. Med.*, 26(11):2389–2430, 2007.

M. Quinn and E. Allen. Changes in incidence of and mortality from breast cancer in England and Wales since introduction of screening. *BMJ: British Medical Journal*, 311(7017):pp. 1391–1395, 1995.

G. K. Reeves, V. Beral, D. Bull, and M. Quinn. Estimating relative survival among people registered with cancer in England and Wales. *Br J Cancer*, 79(1):18–22, Jan 1999.

C. Robertson and P. Boyle. Age, period and cohort models: The use of individual records. *Statist. Med.*, 5(5):527–538, 1986.

C. Robertson, S. Gandini, and P. Boyle. Age-period-cohort models: A comparative study of available methodologies. *Journal of Clinical Epidemiology*, 52:569–583, 1999.

P. S. Rosenberg and W. F. Anderson. Age-period-cohort models in cancer surveillance research: Ready for prime time? *Cancer Epidemiology Biomarkers & Prevention*, 20(7):1263–1268, July 2011.

K. Rostgaard, M. Væth, H. Holst, M. Madsen, and E. Lynge. Age-period-cohort modelling of breast cancer incidence in the Nordic countries. *Statistics in Medicine*, 20:47–61, 2001.

P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 43(3):429–467, Jan. 1994.

P. Royston and M. K. B. Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002.

M. J. Rutherford, P. C. Lambert, and J. R. Thompson. Age-period-cohort modeling. *Stata Journal*, 10(4):606–627, 2010.

M. J. Rutherford, P. W. Dickman, and P. C. Lambert. Comparison of methods for calculating relative survival in population-based studies, June 2011a.

M. J. Rutherford, J. R. Thompson, and P. C. Lambert. Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines. *Statistics in Medicine (submitted)*, 2011b.

C. Sala, E. Morignat, C. Ducrot, and D. Calavas. Modelling the trend of bovine spongiform encephalopathy prevalence in France: Use of restricted cubic spline regression in age-period-cohort models to estimate the efficiency of control measures. *Preventive Veterinary Medicine*, 90(1-2):90 – 101, 2009.

D. Sarfati, T. Blakely, and N. Pearce. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *International Journal of Epidemiology*, 39(2):598–610, Apr. 2010.

W. A. Satariano, K. E. Ragland, and S. K. Van Den Eeden. Cause of death in men diagnosed with prostate carcinoma. *Cancer*, 83(6):1180–1188, Sept. 1998.

W. Sauerbrei, C. Meier-Hirmer, A. Benner, and P. Royston. Multivariable regression model building by using fractional polynomials: Description of SAS, Stata and R programs. *Computational Statistics & Data Analysis*, 50(12):3464 – 3485, 2006.

D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69 (1):239–241, Apr. 1982.

J. Seppänen, S. Heinävaara, and T. Hakulinen. Influence of alternative mammographic screening scenarios on breast cancer incidence predictions (Finland). *Cancer Causes and Control*, 17: 1135–1144, 2006. 10.1007/s10552-006-0055-1.

A. Shah, C. A. Stiller, M. G. Kenward, T. Vincent, T. O. B. Eden, and M. P. Coleman. Childhood leukaemia: long-term excess mortality and the proportion 'cured'. *Br J Cancer*, 99(1):219–223, 2008.

K. Shibuya, M. Inoue, and A. D. Lopez. Statistical modeling and projections of lung cancer mortality in 4 industrialized countries. *Int J Cancer*, 117(3):476–485, Nov 2005.

A. Simonetti, A. Gigli, R. Capocaccia, and A. Mariotto. Estimating complete prevalence of cancers diagnosed in childhood. *Statist. Med.*, 27(7):990–1007, 2008.

B. D. Smith, G. L. Smith, A. Hurria, G. N. Hortobagyi, and T. A. Buchholz. Future of cancer incidence in the United States: burdens upon an aging, changing nation. *J Clin Oncol*, 27: 2758–2765, 2009.

L. K. Smith, P. C. Lambert, J. L. Botha, and D. R. Jones. Providing more up-to-date estimates of patient survival: a comparison of standard survival analysis with period analysis using life-table methods and proportional hazards models. *J Clin Epidemiol*, 57(1):14–20, Jan 2004.

R. Sposto. Cure model analysis in cancer: an application to data from the children's cancer group. *Statist. Med.*, 21(2):293–312, 2002.

StataCorp. 2009. Stata Statistical Software: Release 11. College Station, TX: StataCorp LP.

Statistics Finland. http://www.stat.fi/index_en.html.

E. Svensson, B. Mller, S. Tretli, L. Barlow, G. Engholm, E. Pukkala, M. Rahu, L. Tryggvadttir, F. Langmark, and T. Grotmol. Early life events and later risk of colorectal cancer: age-period-cohort modelling in the Nordic countries and Estonia. *Cancer Causes Control*, 16(3): 215–223, Apr 2005.

E. Sverdrup. Statistiske metoder ved dødelikhetsundersøkelser. statistical memoirs. (in Norwegian), 1967.

N. Tabata, Y. Ohno, R. Matsui, H. Sugiyama, Y. Ito, H. Tsukuma, and A. Oshima. Partial cancer prevalence in Japan up to 2020: Estimates based on incidence and survival data from population-based cancer registries. *Japanese Journal of Clinical Oncology*, 38(2):146–157, Feb. 2008.

M. Talbäck, M. Rosén, M. Stenbeck, and P. W. Dickman. Cancer patient survival in Sweden at the beginning of the third millennium - predictions using period analysis. *Cancer Causes and Control*, 15(9):967–976, Nov. 2004.

L. Teppo, E. Pukkala, and M. Lehtonen. Data quality and quality control of a population-based cancer registry. experience in Finland. *Acta Oncologica (Stockholm, Sweden)*, 33(4):365–369, 1994. PMID: 8018367.

L. Teppo, P. W. Dickman, T. Hakulinen, T. Luostarinen, E. Pukkala, R. Sankila, and B. Söderman. Cancer patient survival - patterns, comparisons, trends: A population-based cancer registry study in Finland. *Acta Oncol*, 38(3):283–294, Jan. 1999.

Thames Cancer Registry. http://www.thames-cancer-reg.org.uk/.

C. Theisen. Predicting the future: projections help researchers allocate resources. *J Natl Cancer Inst*, 95(12):846–848, Jun 2003.

M. Tobias, W. C. Chan, C. Wright, R. Jackson, S. Mann, and L.-C. Yeh. Can the incidence and prevalence of coronary heart disease be determined from routinely collected national data? population-based estimates for New Zealand in 2001-03. *Australian and New Zealand Journal of Public Health*, 32(1):24–27, 2008.

Y.-K. Tu, G. Smith, and M. Gilthorpe. A new approach to age-period-cohort analysis using partial least squares regression: The trend in blood pressure in the glasgow alumni cohort. *PLoS ONE*, 6(4):–, 2011.

A. Verdecchia, R. De Angelis, R. Capocaccia, M. Sant, A. Micheli, G. Gatta, and F. Berrino. The cure for colon cancer: Results from the eurocare study. *International Journal of Cancer*, 77(3):322–329, 1998.

A. Verdecchia, G. D. Angelis, and R. Capocaccia. Estimation and projections of cancer prevalence from cancer registry data. *Statistics in Medicine*, 21(22):3511–3526, 2002.

D. L. Weakliem. A critique of the Bayesian Information Criterion for model selection. *Sociological Methods & Research*, 27(3):359–397, 1999.

J. D. Wessler, N. Pashayan, D. C. Greenberg, and S. W. Duffy. Age-period-cohort analysis of colorectal cancer in East Anglia, 1971-2005. *Cancer Epidemiology*, 34(3):232 – 237, 2010.

WHO. http://www.who.int/cancer/en/.

I. O. L. Wong, B. J. Cowling, C. M. Schooling, and G. M. Leung. Age-period-cohort projections of breast cancer incidence in a rapidly transitioning Chinese population. *Int J Cancer*, 121 (7):1556–1563, Oct 2007.

L. M. Woods, B. Rachet, P. C. Lambert, and M. P. Coleman. 'Cure' from breast cancer among two populations of women followed for 23 years after diagnosis. *Annals of Oncology*, 20(8): 1331–1336, 2009.

S.-H. Xie, J. Gong, N.-N. Yang, L.-A. Tse, Y.-Q. Yan, and I. T.-S. Yu. Time trends and age-period-cohort analyses on incidence rates of nasopharyngeal carcinoma during 1993-2007 in Wuhan, China. *Cancer Epidemiology*, In Press, Corrected Proof:–, 2011.

Y. Yang, W. Fu, and K. Land. A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology*, 34: 75–110, 2004.

F. Yoshimoto, T. Harada, and Y. Yoshimoto. Data fitting with a spline using a real-coded genetic algorithm. *Computer-Aided Design*, 35(8):751 – 760, 2003. Genetic Algorithms.

X. Q. Yu, D. P. Smith, M. S. Clements, M. I. Patel, B. McHugh, and D. L. O'Connell. Projecting prevalence by stage of care for prostate cancer and estimating future health service needs: protocol for a modelling study. *BMJ Open*, pages –, Apr. 2011.

T. Zheng, S. T. Mayne, T. R. Holford, P. Boyle, W. Liu, Y. Chen, M. Mador, and J. Flannery. Time trend and age-period-cohort effects on incidence of esophageal cancer in Connecticut, 1935-89. *Cancer Causes and Control*, 3:481–492, 1992. 10.1007/BF00051361.