The International Journal of Biostatistics

Volume 8, *Issue* 1 2012 *Article* 33

Projecting Cancer Incidence using Age-period-cohort Models Incorporating Restricted Cubic Splines

Mark J. Rutherford, *University of Leicester*
John R. Thompson, *University of Leicester*
Paul C. Lambert, *University of Leicester and Karolinska Institutet*

Recommended Citation:
Rutherford, Mark J.; Thompson, John R.; and Lambert, Paul C. (2012) "Projecting Cancer Incidence using Age-period-cohort Models Incorporating Restricted Cubic Splines," *The International Journal of Biostatistics*: Vol. 8: Iss. 1, Article 33.
DOI: 10.1515/1557-4679.1411

Projecting Cancer Incidence using Age-period-cohort Models Incorporating Restricted Cubic Splines

Mark J. Rutherford, John R. Thompson, and Paul C. Lambert

Abstract

Age-period-cohort models provide a useful method for modeling incidence and mortality rates. There is great interest in estimating the rates of disease at given future time-points in order that plans can be made for the provision of the required future services. In the setting of using age-period-cohort models incorporating restricted cubic splines, a new technique for projecting incidence is proposed. The new technique projects the period and cohort terms linearly from 10 years within the range of the available data in order to give projections that are based on recent trends. The method is validated via a comparison with existing methods in the setting of Finnish cancer registry data. The reasons for the improvements seen for the newly proposed method are twofold. Firstly, improvements are seen due to the finer splitting of the timescale to give a more continuous estimate of the incidence rate. Secondly, the new method uses more recent trends to dictate the future projections than previously proposed methods.

KEYWORDS: age-period-cohort models, incidence models, projecting incidence, cancer incidence

Author Notes: We would like to thank the reviewers for their helpful comments on the manuscript. Mark J Rutherford is funded by a Cancer Research UK Postdoctoral Fellowship (CRUK_A13275).

1 Introduction

Incidence data of the type collected by cancer registries must be projected appropriately if we are to obtain an accurate estimate of the future burden of cancer. This is usually achieved by using Age-Period-Cohort (APC) models that represent the incidence of disease as a product of three components, one based on current age, one on the current time period and the other on the cohort or year of birth (Carstensen (2007)). Unfortunately, the use of these models is not straightforward as they suffer from an identifiability problem due to the linear relationship between age, period and cohort (Holford (1983)). Here we investigate the use of restricted cubic splines for the three terms in the model and show how the fact that restricted cubic splines are linear beyond the final knot can be used to provide good projections of cancer incidence.

Cancer incidence is usually reported in five-year intervals of age and period and consequently it is common practice to fit the age, period and cohort terms in an APC as factors (Zheng et al. (1992), Bergstrom et al. (1996), Gordon et al. (2011), Lee et al. (2011)). However, it has been shown that by using methods of smoothing, such as splines (Durrelman and Simon (1989)), it is possible to model the age, period and cohort terms as varying continuously and so to obtain better fitting models (Carstensen (2007), Heuer (1997), Clements et al. (2005), Holford et al. (2006), Mistry et al. (2011)). Despite this, the use of factors is still wide-spread in applied research.

Many methods have been proposed for making projections from APC models (Knorr-Held and Rainer (2001), Clements et al. (2005), Bray and Møller (2006)) and a large number of them have been used in applied studies (Bray et al. (2001), Rostgaard et al. (2001), Cleries et al. (2009), Lee et al. (2011), Mistry et al. (2011)). Møller et al. (2003) compared fifteen of these methods using data from the Nordic countries. They found that multiplicative APC models tend to over-estimate future incidence and they observed that linear projections need to be tempered or dampened when making long-term prediction. Accordingly they advocated the use of an APC with a power link function together with a linear combination of age, period and cohort terms.

We apply our method based on restricted cubic splines to data from the Finnish Cancer registry on four of the most common types of cancer (breast, lung, colon and pancreas). In particular we investigate the quality of the projections and compare them with those obtained by more standard methods.

2 Description of the Data

The data used in the illustrations is taken from the Finnish Cancer register. Lung, colon and pancreatic cancer were analysed separately for males and for females, while breast cancer is analysed only in females. The incidence data have been restricted to the population aged between 20 years and 80 years and cover the period between 1957 and the end of 2007. In order to assess the quality of the projections, the models were fitted to data from the early years and the resulting projections were compared with the actual data for the later years. Thus to assess the quality of 20 year predictions the model was fitted to data from 1957 to 1987 and the projections were compared with the remaining data.

Cancer incidence is based on a combination of number of cases reported by the cancer registry and data on the corresponding population size taken from Statistics Finland (2012). In practice estimates of the future number of cases of cancer need to project both the incidence and the population size. However, since our examples project over a past period we are able to use exact population sizes for both the fitting of the model and the period of projection and so concentrate on the accuracy of the APC model.

The Finnish registry reports data that can be categorised into one year age groups and one year periods. These can be thought of as placing a rectangular grid over the Lexis diagram that traces an individual's history in age and time. Carstensen (2007) has shown how accurate cohort information can be derived from grouped data by dividing this grid into triangles based on an approach suggested by Sverdrup (1967) and we use this method in all of our examples.

3 Background

3.1 Age-period-cohort models

If the true incidence of disease for people of age a in period p is $\ln\{\lambda(a,p)\}$, then the usual multiplicative APC model can be written using a log-link as:

$$\ln\{\lambda(a,p)\} = f(a) + g(p) + h(c), \tag{1}$$

where $c = p - a$ represents the cohort and f, g and h are functions chosen to represent the pattern in the data.

To overcome the problem of non-identifiability, constraints must be placed on the functions, f, g and h. The most common way of doing this is

to place the constraints on the period and cohort terms after first extracting a linear term that is referred to as drift. We adopt this approach and will often express the drift as part of the period term. Carstensen (2007) provides a full discussion of this and other possible parameterisations.

The required constraints are applied by detrending the period and cohort functions, where the detrended functions will be denoted by $\tilde{g}(p)$ and $\tilde{h}(c)$ respectively. In practice, obtaining the detrended functions that are 0 on average and have no overall trend is achieved by projecting the relevant columns of the design matrix onto the orthogonal complement of the space spanned by the constant and linear term (Holford (1983), Carstensen (2007)). Carstensen (2007) shows that the drift depends on the way in which orthogonality is defined and highlights the difference between using the usual inner product, and an inner product with a defined weight. For the analyses conducted in this paper, the drift is extracted using weights that are proportional to the number of cases for each combination of age, period and cohort.

If both the cohort and period effects are detrended it is possible to fully extract the drift term (δ) as a separate parameter and to write the model in either of the forms:

$$\begin{aligned} \ln\{\lambda(a,p)\} &= f_p(a) + \delta p + \tilde{g}(p) + \tilde{h}(c) \\ &= f_c(a) + \delta c + \tilde{g}(p) + \tilde{h}(c), \end{aligned} \tag{2}$$

where $f_p(a)$ and $f_c(a)$ vary depending on whether the age-specific rates are given relative to period or cohort. Given that $c = p - a$ it follows that:

$$f_p(a) = f_c(a) - \delta a \tag{3}$$

where $\tilde{g}$ and $\tilde{h}$ are the same under either parameterisation.

3.2 Using Restricted Cubic Splines in an APC setting

Factor models assign different levels to each grouped age, period and cohort in equation (1) but an alternative is to use smoothing functions, such as splines. A spline is 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 linear.

Restricted *cubic* splines refer to restricted splines that use cubic polynomials between the knots. They have been used in many forms of regression analysis (Durrelman and Simon (1989)). Cubic polynomials offer sufficient flexibility to capture the shape of most data, provided that appropriate knots are chosen. A restricted cubic spline function can be written in terms of $K-1$ basis functions, where K is the number of knots. The degrees of freedom for the spline function is therefore $K-1$. For knots $k_1, \ldots, k_K$, the spline function $S(x)$ of a given covariate x can be written as:

$$S(x) = \gamma_0 + \sum_{i=1}^{K-1} \gamma_i B_i(x), \tag{4}$$

where $B_1(x) = x$ and, for $i = 2, \ldots, K-1$

$$B_i(x) = (x - k_i)_+^3 - \alpha_i (x - k_1)_+^3 - (1 - \alpha_i)(x - k_K)_+^3 \tag{5}$$

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

$$\alpha_i = \frac{k_K - k_i}{k_K - k_1}. \tag{6}$$

To avoid high levels of correlation between the basis vectors, it is usual to orthoganalise the spline basis; here we use Gram-Schmidt orthogonalization (Golub and van Loan (1996)) and place the knots at the quantiles of the data. The quantiles are defined by placing the knots equally according to the number of events. We use restricted cubic splines for each of the three components of the age-period-cohort models and to ensure identifiability the spline functions are constrained by extracting the linear trend (the drift) from the period and cohort terms, which effectively fixes the slope of both curves. In plots this linear drift is shown as part of the period effect.

4 Description of Methods

4.1 "Spline Drift" Projection

One option for providing projections of cancer incidence from the model defined in equation (2) is to project forward the overall drift parameter, δ, in order to provide future estimates of cancer incidence. This drift parameter is defined by the overall linear trend over the entire follow-up for the range of observed data. This method is advocated by Møller et al. (2003) as the

simplest method of using the Age-Period-Cohort models to provide a linear prediction into the future. The future non-linear period and cohort terms are set equal to the last estimated effect in the range of the data and the future age effects are assumed to be the same as those estimated from the existing data. These terms can then be combined to give unique estimates of the future cancer incidence. The estimate of the linear drift is equivalent under any given parameterisation (Holford (1983)) and consequently the future incidence estimates are invariant under the choice of parameterisation (see Section 4.3).

4.2 "Spline Restriction" Projection

When g and h in equation (1) are modelled by restricted cubic splines the functions are forced to be linear beyond the final knot. This constraint can be used to project the two functions so as to provide model-based projections of incidence. In this case the linear extension of the period and cohort effects for projection will be determined by the latter part of the observed data. The model fitted to the available data is the same as that derived for the standard drift projection approach (as detailed in equation (2)), with an equivalent estimate of the overall drift parameter.

In order to make the projection less dependent on a local trend at the end of the observed data, the final boundary knot can be moved within the range of the observed data to enforce a linear trend to occur from an earlier point in time. In the analyses carried out in this paper we will bring the boundary knot for period and cohort 10 years from the end of the observed data (shown by the dashed vertical lines in Figure 1). The remaining knots are placed at equally-spaced quantiles of the number of events across both period and cohort with the shortened range. For example, for the observed values for the period term (1953-1987), the boundary knot for the restriction model would be placed at 1977 and the knots would then be equally spaced across the range 1953-1977. Figure 1 highlights how the linear constraint can then be used to project the cohort and period terms into the future to provide future estimates of cancer incidence. In Figure 1, the drift term has been included with the period term, which forces the age-effects to have a cross-sectional intepretation. Equivalent projections are obtained if the drift term is allocated instead to the cohort term; these two parameterisations can be seen in equation (2). Allocating the drift to the cohort term will mean that the effect of age should be interpreted longitudinally relative to a reference cohort point.

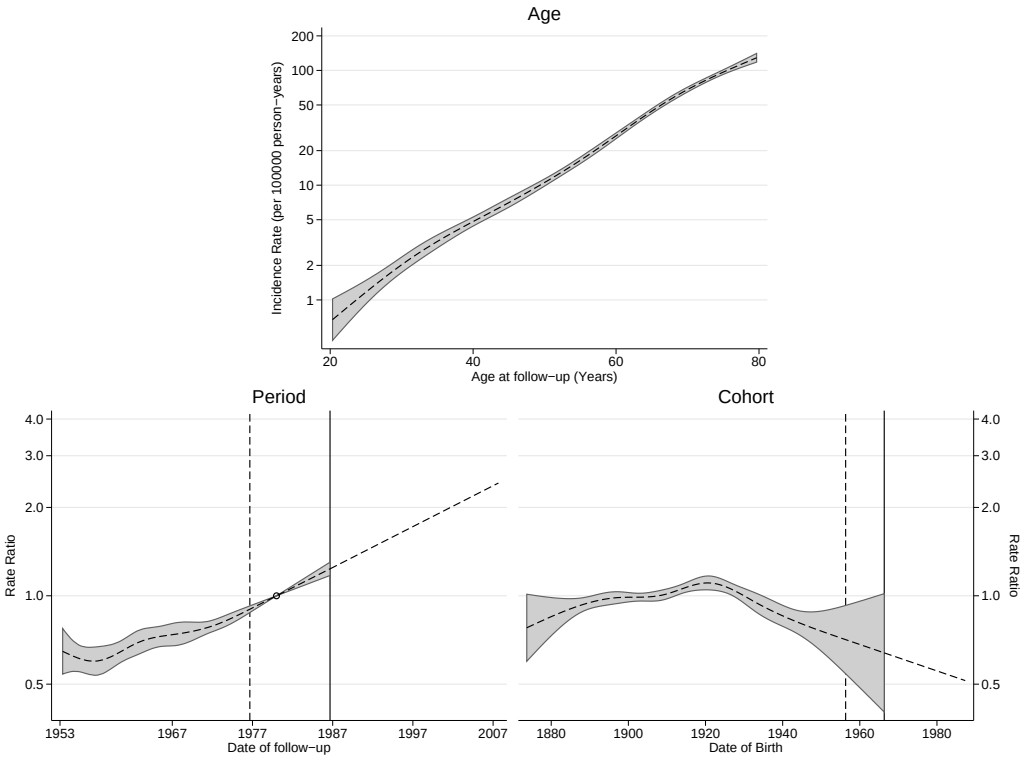

Figure 1: Example of the graphical representation of the age-period-cohort model using restricted cubic splines. The data used are for the incidence of Finnish colon cancer for males. The drift term is attributed to the period curve, and the age curves are the fitted rates in the reference period (1980 - indicated by the hollow circle).

4.3 Projection Invariance

It is vital that any projection estimate provides unique estimates of future cancer incidence irrespective of the chosen parameterisation of the model. In order for the projections to be unique, the projection technique must provide equivalent projections when an arbitrary linear trend is assigned to any of the given terms. This is satisfied if the future period and cohort terms are linear extensions of the fitted functions (Osmond (1985)).

Consider the "spline drift" method described in Section 4.1 applied to a future year p^* and a given value of age, a^*. Let $p_{\max}$ be the maximum observed value for period and $a_{\min}$ be the youngest observed age, then we

have $c_{\max} = p_{\max} - a_{\min}$ as the maximum observed value for cohort. If we have $p^* - a^*(= c^*) > c_{\max}$ then it is necessary to consider the projection of both g and h in order to evaluate the value of $\ln\{\lambda(a^*, p^*)\}$. For the future projections the functions $\tilde{g}$ and $\tilde{h}$ are evaluated at the last observed value, and the drift term is used to project forward to the future period:

$$\ln\{\lambda(a^*, p^*)\} = f_p(a^*) + \delta p^* + \tilde{g}(p_{\max}) + \tilde{h}\,(p_{\max} - a_{\min})\,. \tag{7}$$

The fact that the drift term can be allocated to the function $\tilde{g}$ to give g simply involves a projection that is then of a slope δ from $g(p_{\max})$ (similarly for the reallocation of the drift to the function $\tilde{h}$). Therefore, we have invariance under a reallocation of a linear component.

Alternatively, if $p^* - a^* < c_{\max}$ then the cohort term can be evaluated from within the range of the data and the projection will instead be given by:

$$\ln\{\lambda(a^*, p^*)\} = f_p(a^*) + \delta p^* + \tilde{g}(p_{\max}) + \tilde{h}(p^* - a^*). \tag{8}$$

Using a similar argument to the one given above, the generalisation to other ages which do not involve a projection of h also follows.

To satisfy the property of invariance under projection for the spline restriction approach (Section 4.2), it is again sufficient to show that the projection of the functions would be equivalent under a reallocation of a linear component. Similarly to the justification for the spline drift approach, consider a future year p^* and a given age a^*. In the case where we have $p^* - a^* > c_{\max}$ for the spline restriction approach, we can write the projected estimate as:

$$\begin{aligned}\ln\{\lambda(a^*, p^*)\} = f_p(a^*) & \\ + \delta p^* + \big(\tilde{g}(p_{\max}) + \beta_1(p^*)\big) & \\ + \big(\tilde{h}(c_{\max}) + \beta_2(p^* - a^*)\big) &\end{aligned} \tag{9}$$

where, β_1 is the linear slope of the spline term for period beyond the boundary knot at $p_{\max}$ (or $p_{\max} - 10$ if the boundary knot is moved 10 years within the range of the data) and β_2 is the linear slope of the spline term for cohort beyond the boundary knot at $c_{\max}$ (or $c_{\max} - 10$). For the future projection at (a^*, p^*) the functions $\tilde{g}$ and $\tilde{h}$ are evaluated at the last observed value and the linear function of the two spline terms are projected forward. The fact that the drift term can be allocated to the function $\tilde{g}$ to give g simply involves a projection

that is then of a slope $\delta + \beta_1$ from $g(p_{\max})$, similarly for the reallocation of the drift to the function $\tilde{h}$. Therefore, we have:

$$\begin{aligned}
\ln\{\lambda(a^*, p^*)\} &= f_p(a^*) \\
&\quad + \delta p^* + \big(\tilde{g}(p_{\max}) + \beta_1 p^*\big) \\
&\quad + \big(\tilde{h}(c_{\max}) + \beta_2(p^* - a^*)\big) \\
&= f_p(a^*) \\
&\quad + \big(g(p_{\max}) + (\delta + \beta_1)p^*\big) \\
&\quad + \big(\tilde{h}(c_{\max}) + \beta_2(p^* - a^*)\big) \\
&\text{using equation (3)} \\
&= f_c(a^*) \\
&\quad + \big(\tilde{g}(p_{\max}) + \beta_1 p^*\big) \\
&\quad + \big(h(c_{\max}) + (\delta + \beta_2)(p^* - a^*)\big),
\end{aligned} \tag{10}$$

and consequently invariance under a reallocation of a linear component. For the case where $p^* - a^* < c_{\max}$ a similar argument can be undertaken, and there is no need to project the cohort term in this case.

4.4 Altering the Link Function

The standard link function used for age-period-cohort models is the log, however Engeland et al. (1993) suggested that the exponential growth that this introduces for the projections leads to an overestimation in the projected incidence, particularly for long-term follow-up. Consequently, the authors propose a power link function (with a power of $\frac{1}{5}$) to dampen the exponential growth. The selection of $\frac{1}{5}$ has been evaluated empirically using Nordic data (Engeland et al. (1993), Møller et al. (2003)).

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

$$\{\lambda(a, p)\}^{\frac{1}{5}} = f(a) + g(p) + h(c), \tag{11}$$

and this link can be used in conjunction with either of the projection methods outlined above.

4.5 Comparison of Projection Methods

The spline drift and spline restriction method of projection were used in combination with both the log and the power link functions. All of the analyses were

carried out using a user-written command (Rutherford, Lambert, and Thompson (2010)) for the statistical software package, Stata (StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP (2011)). The program extracts the drift term and fits the model as a generalised linear model (GLM) with an appropriate offset and a Poisson error structure. Both the log link and power link function are applied to the GLM in the comparisons. We used 8 degrees of freedom (7 internal knots, 9 knots overall) for the age and cohort spline terms and 5 degrees of freedom (4 internal knots, 6 including the boundary knots) for the period spline term for each of the modelling approaches. The knots were placed at equal quantiles for the number of events across the relevant variables and at equal quantiles over a restricted range (for period and cohort) for the spline restriction approach. The sensitivity to the selection of knots is investigated in Section 5.5.1.

Long (15-20 years into the future) and short-term (5-10 years into the future) predictions are compared for each of the methods using a similar approach to that undertaken by Møller et al. (2003) when analysing data from the Nordic countries. The analysis was conducted for all ages combined, and the comparison was based on the absolute value of the total relative difference between the total observed and total predicted number of cases, ($|\text{observed} - \text{predicted}| * 100\%/\text{observed}$). Møller et al. (2003) compared the total number of predicted cases over a five-year period to the total number of observed cases over the same period. This tends to mask the benefit of spline approaches that give smoothed estimates for each year. Consequently, we used the same measure of relative difference, but average it over each single year of the prediction window.

Long term projections were made over the period 2003-2007 using observed rates up to the end of 1987. Two different 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 up to the end of 1997. The comparison of the two short-term predictions allows for an assessment of the consistency of the estimation approaches over time.

5 Application

5.1 Short-term Projections

Table 1 A) shows the comparison of the different methods in terms of the average absolute value of the total relative difference for each of the cancer

Table 1: A) Observed data until the end of 1987; 10 year prediction for the period 1993-1997. The values are the average yearly absolute relative difference (%) between the observed and predicted number of cases for all ages combined. B) Observed data until the end of 1997; 10 year prediction for the period 2003-2007.

(A)

Link Function		Log		Power
Cancer Site	Restriction	Drift	Restriction	Drift
Breast (Females)	5.66	2.36	9.93	8.91
Colon (Males)	12.20	4.97	5.64	3.50
Colon (Females)	11.26	13.70	4.05	5.59
Lung (Males)	6.46	14.60	12.29	21.26
Lung (Females)	9.56	7.48	4.62	2.91
Pancreas (Males)	6.23	14.74	6.73	14.80
Pancreas (Females)	5.44	15.27	3.51	11.17
Mean	8.12	10.44	6.68	9.73

(B)

Link Function		Log		Power
Cancer Site	Restriction	Drift	Restriction	Drift
Breast (Females)	8.75	2.60	4.61	4.12
Colon (Males)	5.25	8.43	4.68	4.39
Colon (Females)	8.95	15.77	5.58	8.40
Lung (Males)	9.83	3.20	3.37	15.73
Lung (Females)	10.77	3.83	11.79	7.57
Pancreas (Males)	16.72	10.58	13.98	9.46
Pancreas (Females)	5.44	5.16	5.39	5.32
Mean	9.39	7.08	7.06	7.86

sites for the period from 1993 until the end of 1997. The 'Spline Restriction" approach appears to perform particularly well in the case of male lung cancer, and for pancreatic cancer for both males and females, while in the case of colon cancer for males the drift approach is better. The reasons for the differences in performance are assessed in Figures 2 and 3. The power link is not uniformly better than the standard link.

The reasons for the difference observed in Table 1 A) can be investigated

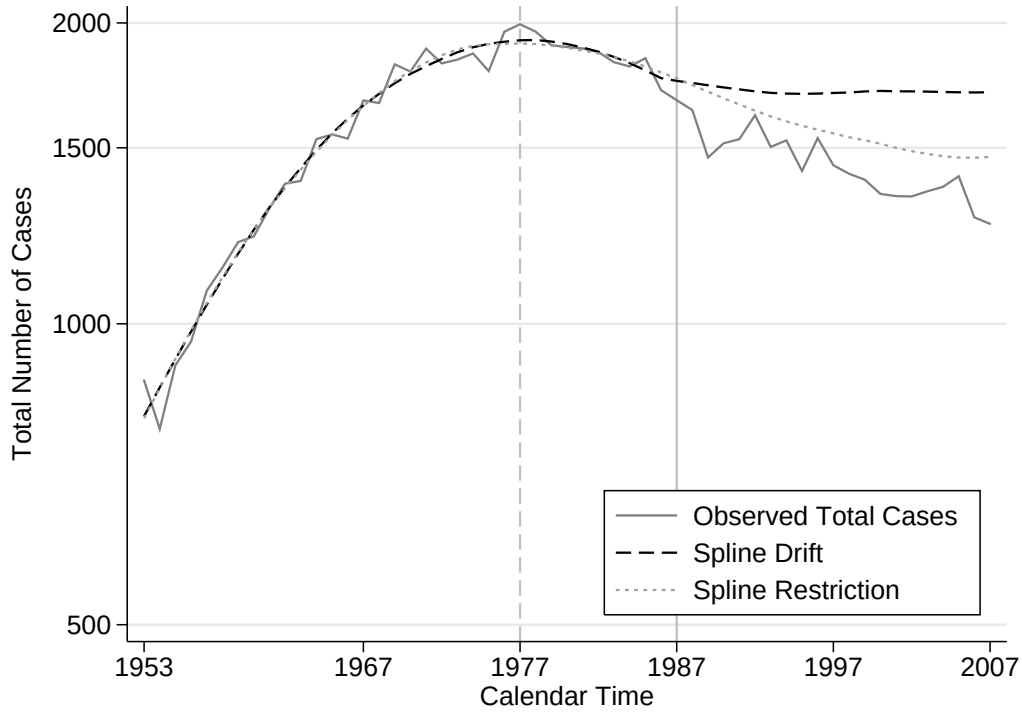

Figure 2: Projections from 1987 for male lung cancer patients for the total number of cases for all ages. GLM fitted with a log link function.

by plotting the historical data together with the projections. Figure 2 gives the graphical representation of the results for the male lung cancer patients using a log-link. There is a clear change in the pattern of the lung cancer cases over time with a substantial decrease in incidence from the 1970s onwards. The method using the spline restriction, which is dominated by the change in the last 10 years, outperforms the drift approach that bases the linear projection on the entire range of the data.

Figure 3 gives the graphical representation of the results for the pancreatic cancer data for females. The figure shows the fitted curves for the two projection methods for the power link function, together with the observed values of the total number of cases. The 10 year window used for the spline restriction shows a smaller gradient than is observed over the longer observation window, consequently the spline restriction approach gives better projections after 1987.

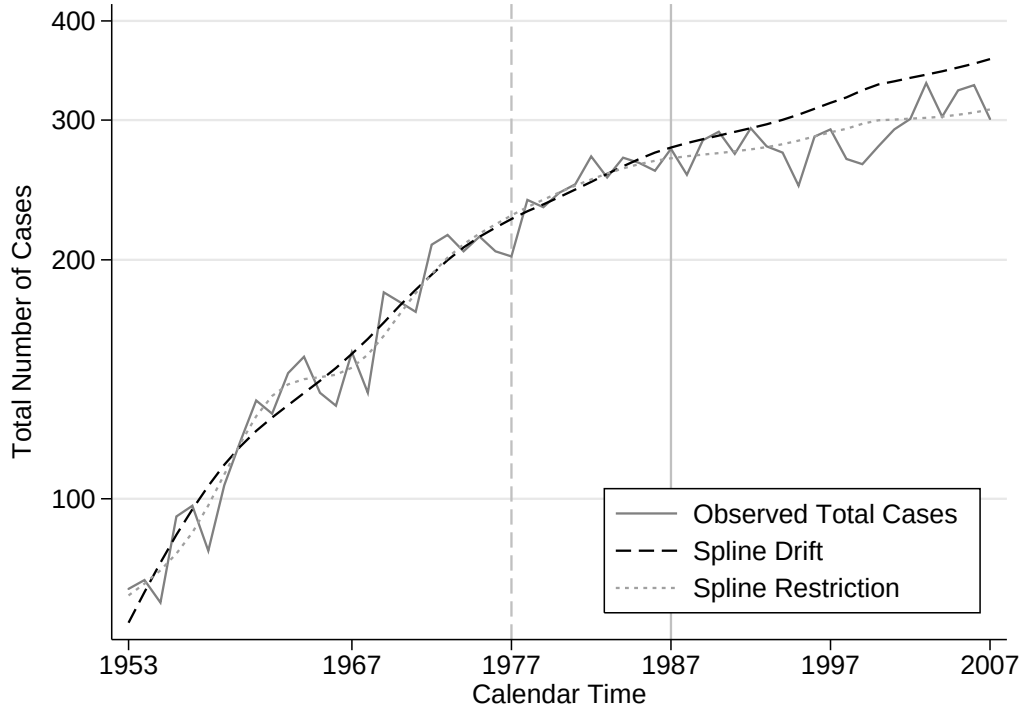

Figure 3: Projections from 1987 for female pancreatic cancer patients for the total number of cases for all ages. GLM fitted with a power link function.

Figure 4 gives the graphical representation of the results for the colon cancer data for males using the power link function. The projection method that uses the drift over the entire observation period performs well as shown in Tables 1 A) and 2. 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.

Table 1 B) contains the projections for data observed until the end of 1997, and projected until the end of the available data in 2007. These are short-term projections but are using a longer range of data than those given in Table 1 A). It is clear from the results contained in Table 1 B) that there is a lack of consistency across time-points for the "best" method of estimation for any given cancer site. This highlights the need for careful consideration

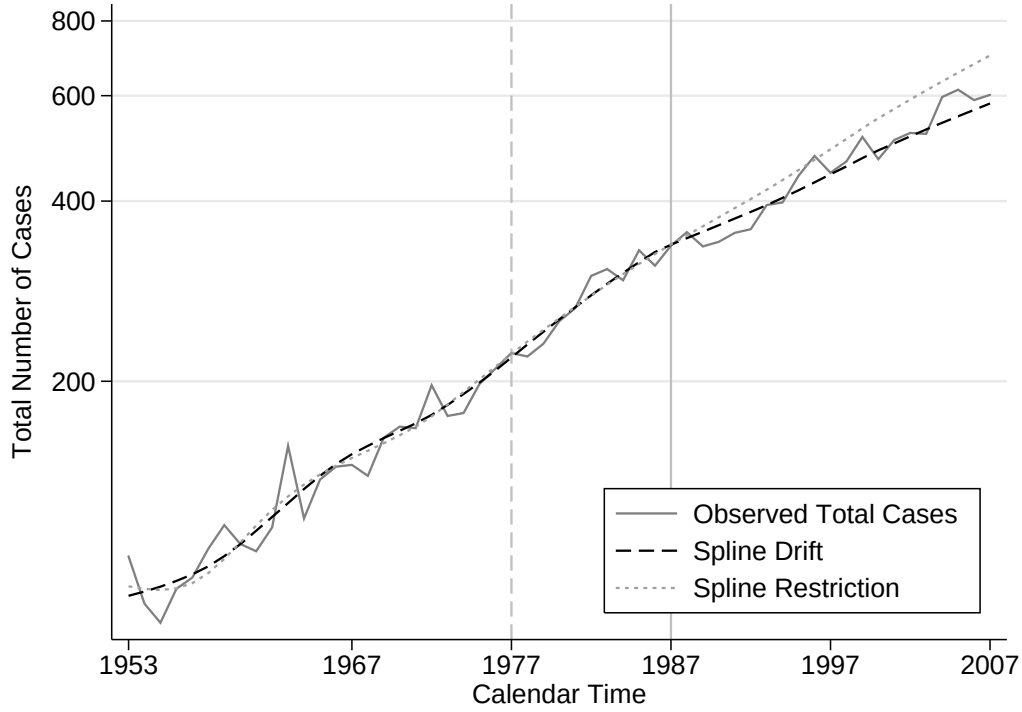

Figure 4: Projections from 1987 for male colon cancer patients for the total number of cases for all ages. GLM fitted with a power link function.

when choosing a method of projection.

5.2 Long-term Projections

Table 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 may well yield better predictions. However, 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. Figures 2, 3 and

Table 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.

Link Function		Log		Power
Cancer Site	Restriction	Drift	Restriction	Drift
Breast (Females)	5.04	2.20	15.50	17.24
Colon (Males)	26.34	6.80	12.48	5.57
Colon (Females)	29.44	38.74	11.50	16.81
Lung (Males)	10.41	28.07	26.50	53.74
Lung (Females)	6.28	5.58	4.44	9.26
Pancreas (Males)	12.11	6.70	7.14	8.31
Pancreas (Females)	4.81	19.65	5.37	9.75
Mean	13.49	15.39	11.85	17.24

4 also show the long-term projections for three of the cancer sites from 1987 onwards.

5.3 Age-specific Projections

Figure 5 shows the results for pancreatic cancer for females split by various age categories. Age is still modelled continuously, but the total number of cases are summed over the selected age categorisations to evaluate the projections by age. The youngest age category has fewer total cases due to the strong association between incidence and age. Therefore, the information for the older age categories dominates the overall shape. However, it is important that the projections at specific ages can also be assessed to evaluate the projection approaches.

5.4 Evaluating Uncertainty

Figure 6 shows the resulting projections with the model-based confidence intervals also shown for both of the projection methods for the lung cancer data. The confidence intervals can be estimated from the model through use of the delta method, which makes use of a Taylor series expansion to obtain the variance-covariance matrix, using the Stata command `predictnl` (StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP (2011)). These confidence intervals are purely based on uncertainty in the

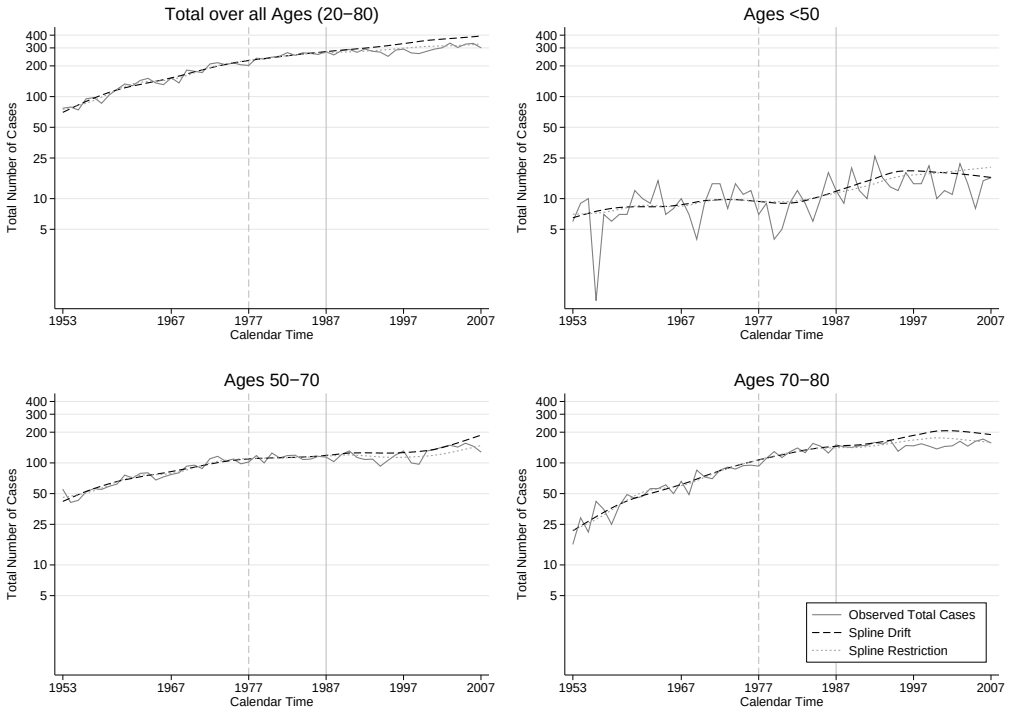

Figure 5: Age-specific projections from 1987 for female pancreatic cancer patients. GLM fitted with a power link function.

estimated parameters for a particular model choice and do not account for the uncertainty that is associated with the choice of assumption for the projection. The plots given in Section 5.5.2 highlight that different projections will be made under a variety of assumptions for the linearity constraint. A combination of these two sources of uncertainty is necessary to fully appreciate the total uncertainty when making projections.

5.5 Sensitivity Analyses

5.5.1 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 3. The analysis

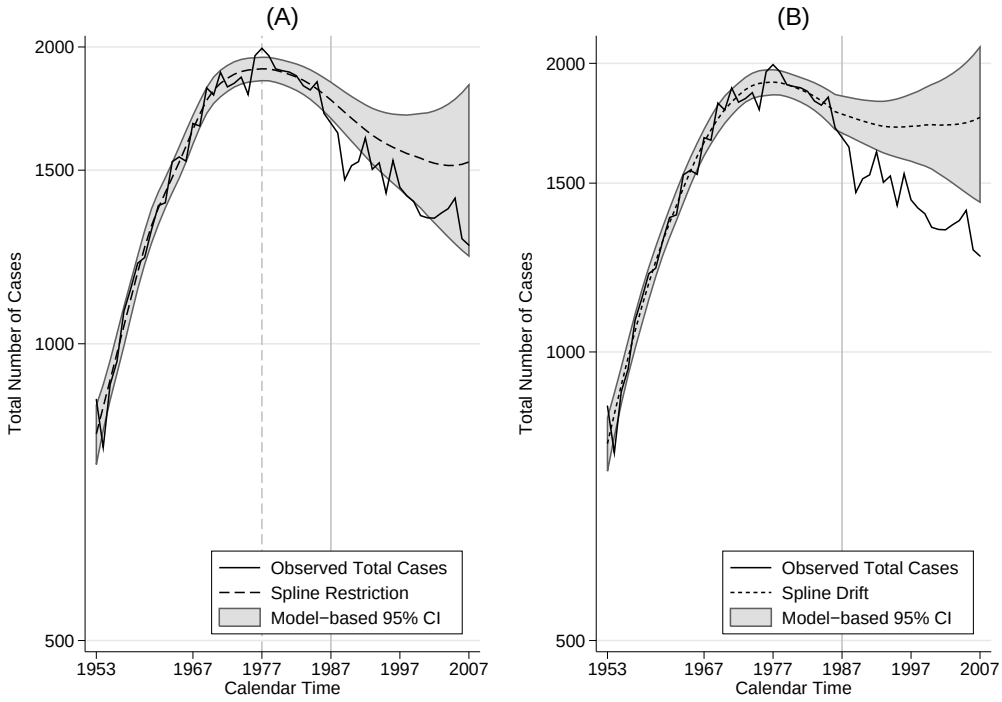

Figure 6: Model-based confidence intervals for projection from 1987 for male lung cancer patients. GLM fitted with a log link function. A) Shows the estimates for the spline drift approach. B) shows the estimates for the spline restriction approach.

for Table 1 A) was performed for each model whilst varying the number of knots for period (degrees of freedom of 3, 5, 6, and 8) keeping 8 degrees of freedom for the age and cohort terms. The degrees of freedom used to carry out the original analysis were 5 degrees of freedom for period, and 8 for the age and cohort terms. The knots were placed at equally spaced quantiles of the relevant variables, and over a shorter time-range for the spline restriction approach as detailed in Section 4.2.

Table 3 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 cancer in males and lung cancer in females quite substantial difference do occur. Smaller differences are also evident for pancreatic cancer in females and colon cancer in females when

a low number of knots are used. In the case of colon cancer for males, we can see that the models with an increasing number of knots for period seem to be giving poorer estimates than the simplest model. Figure 4 shows that there seems to be an almost linear growth in the total number of cases over time suggesting that the models with many knots will overfit the period effect. This is confirmed by looking at the AIC and BIC (values not shown), 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; the simpler models are underfitting the effect of period so that increasing the degrees of freedom leads to better projections. The degrees of freedom preferred for period (from the models compared) differ between the AIC (6 degrees of freedom gives the lowest) and the BIC (5 degrees of freedom gives the lowest). The "best" results for the projection are observed for the model with 8 degrees of freedom for period for the "Spline Restriction" approach. A "better" fitting model to the observed data does not necessarily lead to "the best" projections. In real analyses when projections are actually made into the future rather than into a period where we know the number of cases, it will not be possible to compare the models with different 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.

Table 3: The compared models relate to the different choice of degrees of freedom (df) for Period (3, 5, 6 and 8 degrees of freedom were used respectively for Period, and 8 degrees of freedom were used for Age and Cohort for each model). The values given are equivalent to the values in Table 1 A). 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. 5 degrees of freedom for Period were used in the actual analyses. The knots were placed at equally spaced quantiles of the relevant variables.

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

5.5.2 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 observed data will reduce the weight given to the most recent data 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 paper, 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.

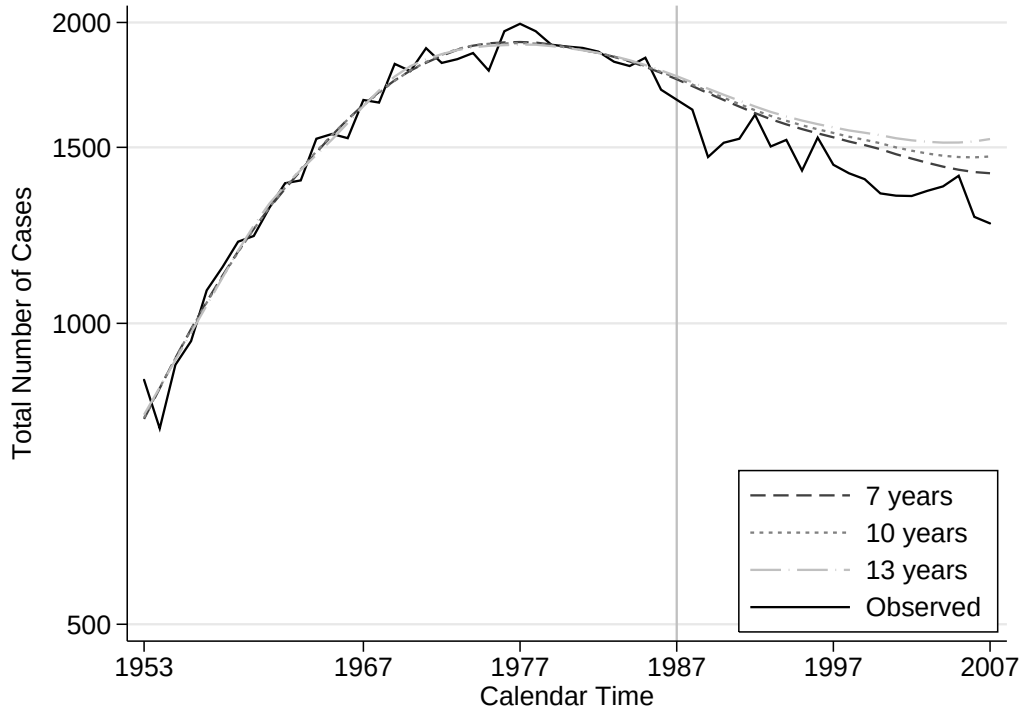

Figure 7: Projections from 1987 for male lung cancer patients for the total number of cases for all ages. The different lines correspond to moving the boundary knot to different points (7, 10 and 13 years) within the range of the data.

Figure 7 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 main analyses. 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.

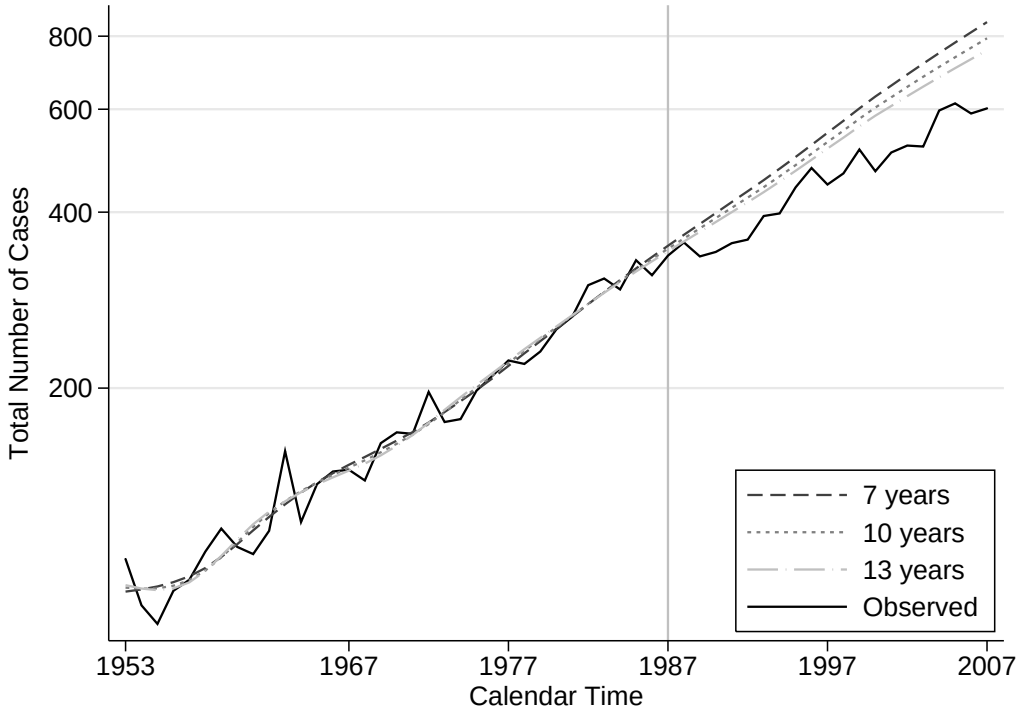

Figure 8: Projections from 1987 for male colon cancer patients for the total number of cases for all ages. The different lines correspond to moving the boundary knot to different points (7, 10 and 13 years) within the range of the data.

Figure 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, 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. Provided that the boundary knot is not placed at the very edge of the available data, each of the projections made will be

largely dictated 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.

6 Discussion

Cancer incidence projections can be made by using the fact that restricted cubic splines are linear beyond the boundary knot. The linear projection beyond the range of the data will be dictated by the shape of the data towards the end of the observation period ensuring that the projections give increased weight to more recent trends than standard approaches. Standard projections using the drift can also be made in the same modelling framework and a direct comparison between the two approaches has been made using a range of cancer sites. The difference in assumptions relates to the "recentness" of the trend that is projected forward.

For the projections made in these analyses, actual population sizes were used as we used historical data. For future predictions of the number of cases, further errors will be introduced by the inaccuracies in the forecasting of population data. 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 known to be accurate.

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 (Verdecchia, Angelis, and Capocaccia (2002), Heinävaara and Hakulinen (2006)). Of course, calculations can be made for the uncertainty for the parameters in a given model, and prediction intervals can be put on the estimated rates (Elkum (2005), Møller, Weedon-Fekjær, and Haldorsen (2005)). However, these intervals can be very narrow in population-based cancer studies where lots of information is available (Møller et al. 2007). The proposed prediction intervals do not take into account the bias introduced by making an untestable assumption about the future rates based on the available data. Empirical evaluations of the prediction intervals have shown that caution should be taken when interpreting them (Møller et al. (2005)). Further to this, there have been claims that these intervals should not be reported because they may be wrongly interpreted (Møller et al. (2007)).

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 Møller et al. (2003). In this paper, we put this 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 (Møller et al. (2003)) can easily be applied to the new approach.

References

Bergström, R., H.-O. Adami, M. Möhner, W. Zatonski, H. Storm, A. Ekbom, S. Tretli, L. Teppo, O. Akre, and T. Hakulinen (1996): "Increase in testicular cancer incidence in six European countries: a birth cohort phenomenon," *Journal of the National Cancer Institute*, 88, 727–733.

Bray, F. and B. Møller (2006): "Predicting the future burden of cancer." *Nat Rev Cancer*, 6, 63–74.

Bray, I., P. Brennan, and P. Boffetta (2001): "Recent trends and future projections of lymphoid neoplasms–a Bayesian age-period-cohort analysis." *Cancer Causes Control*, 12, 813–820.

Carstensen, B. (2007): "Age-period-cohort models for the Lexis diagram," *Statistics in Medicine*, 26, 3018–3045.

Clements, M. S., B. K. Armstrong, and S. H. Moolgavkar (2005): "Lung cancer rate predictions using generalized additive models." *Biostatistics*, 6, 576–589.

Cleries, R., J. M. Martínez, J. Valls, L. Pareja, L. Esteban, R. Gispert, V. Moreno, J. Ribes, and J. M. Borràs (2009): "Life expectancy and age-period-cohort effects: analysis and projections of mortality in Spain between 1977 and 2016." *Public Health*, 123, 156–162.

Durrelman, S. and R. Simon (1989): "Flexible regression models with cubic splines." *Statistics in Medicine*, 8, 551–561.

Elkum, N. B. (2005): "Predicting confidence intervals for the age-period-cohort model," *Journal of Data Science*, 3, 403–414.

Engeland, A., T. Haldorsen, S. Tretli, T. Hakulinen, L. G. Hörte, T. Luostarinen, K. Magnus, G. Schou, H. Sigvaldason, and H. H. Storm (1993): "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.

Golub, G. H. and C. F. van Loan (1996): *Matrix Computations, 3rd Edition*, Johns Hopkins University Press.

Gordon, P., F. Artaud, A. Aouba, F. Laurent, V. Meininger, and A. Elbaz (2011): "Changing mortality for motor neuron disease in France (1968-2007): an age-period-cohort analysis," *European Journal of Epidemiology*, 1–9, 10.1007/s10654-011-9595-0.

Heinävaara, S. and T. Hakulinen (2006): "Predicting the lung cancer burden: accounting for selection of the patients with respect to general population mortality," *Statistics in Medicine*, 25, 2967–2980.

Heuer, C. (1997): "Modeling of time trends and interactions in vital rates using restricted regression splines," *Biometrics*, 53, 161–177.

Holford, T. R. (1983): "The estimation of age, period and cohort effects for vital rates," *Biometrics*, 39, 311–324.

Holford, T. R., K. A. Cronin, A. B. Mariotto, and E. J. Feuer (2006): "Changing patterns in breast cancer incidence trends." *J Natl Cancer Inst Monogr*, 19–25.

Knorr-Held, L. and E. Rainer (2001): "Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction." *Biostatistics*, 2, 109–129.

Lee, T. C. K., C. B. Dean, and R. Semenciw (2011): "Short-term cancer mortality projections: a comparative study of prediction methods." *Statistics in Medicine*, 30, 3387–3402.

Mistry, M., D. M. Parkin, A. S. Ahmad, and P. Sasieni (2011): "Cancer incidence in the united kingdom: projections to the year 2030." *Br J Cancer*, 105, 1795–1803.

Møller, B., H. Fekjær, T. Hakulinen, H. Sigvaldason, H. H. Storm, M. Talbäck, and T. Handorsen (2003): "Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches," *Statistics in Medicine*, 22, 2751–2766.

Møller, B., H. Weedon-Fekjær, and T. Haldorsen (2005): "Empirical evaluation of prediction intervals for cancer incidence," *BMC Medical Research Methodology*, 5, 21.

Møller, H., L. Fairley, V. Coupland, C. Okello, M. Green, D. Forman, B. Møller, and F. Bray (2007): "The future burden of cancer in England: incidence and numbers of new patients in 2020." *Br J Cancer*, 96, 1484–1488.

Osmond, C. (1985): "Using age, period and cohort models to estimate future mortality rates," *International Journal of Epidemiology*, 14, 124–129.

Rostgaard, K., M. Væth, H. Holst, M. Madsen, and E. Lynge (2001): "Age-period-cohort modelling of breast cancer incidence in the Nordic countries," *Statistics in Medicine*, 20, 47–61.

Rutherford, M. J., P. C. Lambert, and J. R. Thompson (2010): "Age-period-cohort modeling," *Stata Journal*, 10, 606–627.

StataCorp. 2011. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP (2011): .

Statistics Finland (2012): "http://www.stat.fi/index_en.html," .

Sverdrup, E. (1967): "Statistiske metoder ved dødelikhetsundersøkelser. statistical memoirs." (in Norwegian).

Verdecchia, A., G. D. Angelis, and R. Capocaccia (2002): "Estimation and projections of cancer prevalence from cancer registry data," *Statistics in Medicine*, 21, 3511–3526.

Zheng, T., S. T. Mayne, T. R. Holford, P. Boyle, W. Liu, Y. Chen, M. Mador, and J. Flannery (1992): "Time trend and age-period-cohort effects on incidence of esophageal cancer in Connecticut, 1935-89," *Cancer Causes and Control*, 3, 481–492.