Predicting Confidence Intervals for the
Age-Period-Cohort Model

Naser B. Elkum

King Faisal Specialist Hospital and Research Centre

Abstract:

Forecasting incidence and/or mortality rates of cancer is of special interest to epidemiologists, health researchers and other planners in predicting the demand for health care. This paper proposes a methodology for developing prediction intervals using forecasts from Poisson APC models. The annual Canadian age-specific prostate cancer mortality rates among males aged 45 years or older for the period between 1950 and 1990 are calculated using 5-year intervals. The data were analyzed by fitting an APC model to the logarithm of the mortality rate. Based on the fit of the 1950 to 1979 data, the known prostate mortality in 1980 to 1990 is estimated. The period effects, for 1970-1979, are extended linearly to estimate the next ten period effects. With the aims of parsimony, scientific validity, and a reasonable fit to existing data two different possible forms are evaluated namely, the age-period and the age-period-cohort models. The asymptotic 95% prediction intervals are based on the standard errors using an assumption of normality (estimate $\pm 1.96\times$ standard error of the estimate).

Key words: Age-period-cohort models, cancer incidence and mortality, projections.

1. Introduction

Forecasting incidence and/or mortality rates of cancer is of special interest to epidemiologists, health researchers and other planners in predicting the demand for health care (Schaubel *et al.* 1998). The simplest way to predict cancer incidence and/or mortality rates is to extrapolate from past trends by fitting a parametric model to the observed cancer mortality cases (MacNeill *et al.* 1995 and Hakulinen and Dyba 1994). However, the point estimates obtained from fitted models, and their associated variability, depend on the parametric form of the model. It is often difficult to choose between different parametric forms, because several models may produce equally good fits to the data but offer very different predictions (Dyba and Hakulinen 2000). Using linear modeling, a "prediction interval" was developed by Hakulinen and Dyba (1994), by assuming a

Poisson distribution for the number of incident cases, but such linear trends are not likely to last indefinitely and both the year in which the disease was diagnosed (period); and the year in which the subject was born (cohort) may contribute simultaneously to the observed rates of cancer incidence and mortality.

This has led to the increasing use of log-linear Poisson age-period-cohort (APC) models for the statistical analysis of this type of data (Osmond 1985). The age of a subject at the time of their diagnosis (age), along with period and cohort, are the three time factors typically used in studying patterns of morbidity and mortality rates (Osmond and Gardner 1983 and Holford 1985). However, although these have received much attention in the literature (Kupper *et al.* 1983 and Holford 1991), their complex inter-relationship does pose identification problems, since knowledge of any two of these factors automatically determines the effect of the third (Kupper *et al.* 1985 and Holford 1983). Fortunately, projections based on the APC model are uniquely determined (Holford 1985). APC models were used to forecast the incidence and mortality of lung cancer in England, Wales and Korea, and the mortality due to all cancers in Switzerland (Osmond 1985, Negri *et al.* 1990, and Jee *et al.* 1998). A Bayesian APC model has also been used to predict incidence of Hodgkin's disease in Oxford (Bray 2002). In reviewing these methods, Bray (2002) showed that one particular model (the Osmond model 1985) always provides the best fit to the data. However, all these models are only capable of providing point estimates of future rates; the calculation of prediction confidence intervals for the Poisson regression models has remained a hard task.

A more detailed analysis of these shortcomings inherent in the earlier approaches, has stimulated us to investigate the advantages of applying Poisson APC models in obtaining more accurate predictions and their associated confidence limits. Our specific objective here is to develop prediction intervals using forecasts from Poisson APC models.

2. The APC Model

Descriptive epidemiologists are interested in the presentation and interpretation of temporal variation in cancer rates. The issue here, in its simplest form, concerns the analysis of a set of rates arranged in a two-way table by age group and calendar period. A specific model commonly applied to this type of cross-classified data is APC model. It expresses each cell of the mortality table in the general form:

$$\lambda_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k \tag{2.1}$$

where α_i $(i = 1, \ldots, I)$ represents the i-th age group effect; β_j $(j = 1, \ldots, J)$ represents the j-th calendar period effect; γ_k $(k = 1, \ldots, I+J-1)$ represents the

k-th cohort effect, and the dependent variable $\lambda_{ijk} = f(n_{ijk}/N_{ijk})$ is a function of the mortality rate; n_{ijk} is the observed number of deaths in the i-th age group, j-th calendar period and k-th cohort; N_{ijk} is the corresponding number of subjects at risk.

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j + \sum_{k=1}^{I-J-1} \gamma_k = 0.$$

These restrictions imply, for example, that only the first $(I-1)$ age effects, the first $(J-1)$ period effects, and the first $(I+J-2)$ cohort effects in model (2.1) require estimation.

3. Estimation

Model (2.1) belongs to a class of generalized linear models and can be fitted using standard methods developed for those models (McCullagh and Nelder 1989). For this class of models, a function of the mean is linearly related to a set of regression variables. Under Poisson error, the model is linear in the logarithm of the expected cell count $\mu_{ijk} = E(n_{ijk})$. In our case, the expected value in any cell is given by

$$\mu_{ijk} = E(n_{ijk}) = N_{ijk} \exp\{\mu + \alpha_i + \beta_j + \gamma_k\} = N_{ijk} \exp\{\eta_{ijk}\},$$

$i = 1, \ldots, I; j = 1, \ldots, J; k = 1, \ldots, K.$
The actual model being fitted is:

$$\log(E(n_{ijk})) = \mu + \alpha_i + \beta_j + \gamma_k + \log(N_{ijk}) = XB \tag{3.1}$$

where $\log(N_{ijk})$ is referred to as an offset and is treated as a known constant, X is the model matrix and B is the column vector of parameters μ, α, β and γ. Let b be the estimate of B, then the maximum likelihood estimates are obtained iteratively from the equations

$$(X^T \Gamma X)^{(m-1)} b^{(m)} = (X^T \Gamma Z)^{(m-1)} \tag{3.2}$$

where the superscript $(m-1)$ denotes evaluation at $b^{(m-1)}$, Γ the diagonal matrix with elements $\Gamma_{ijk}^{-1} = var(\eta_{ijk})(\partial \eta_{ijk}/\partial \mu_{ijk})^2$ and $z_{ijk} = \eta_{ijk} + (n_{ijk} - \mu_{ijk})(\partial \eta_{ijk}/\partial \mu_{ijk})$. The procedure begin by using some initial approximation $b^{(0)}$ to evaluate Z and Γ, then (3.2) is solved to give $b^{(1)}$ which in turn is used to obtain better approximations for Z and Γ, and so until adequate convergence is achieved. When the difference between successive approximations $b^{(m-1)}$ and $b^{(m)}$ is sufficiently small, $b^{(m)}$ is taken as the maximum likelihood estimate.

4. Variance of Future Estimates for Age-Period-Cohort Model

The motivation for using an age, period and cohort approach to the estimation of future mortality rates is that it takes account of both period and cohort effects, in addition to age effects. Let

$$\underline{\psi}^T = (\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_1, \ldots, \gamma_K) \tag{4.1}$$

The first step is to obtain $\hat{\alpha}_i, \hat{\beta}_j$ and $\hat{\gamma}_k$ the estimates of α_i, β_j and γ_k produced by iteratively reweighted least squares. To extrapolate, keep the α's unchanged; that is, no extension to other ages is required. The estimates for future period values and cohort values will be obtained by linear regression applied to the most recent period and cohort values.

Let $\hat{\theta}_{ilk}$ be the maximum likelihood estimate of the expected number of cases in the i-th age group, l-th future year $(I = J+1, \ldots)$ and the corresponding k-th cohort. That is

$$\hat{\theta}_{ilk} = N_{ilk} \exp\{\underline{\alpha}_{ilk}^t \underline{\hat{\psi}}\} \tag{4.2}$$

where the constant N_{ilk} represents the size of the population at risk in the i-th age group, l-th future period and k-th future cohort, and $\underline{\alpha}_{ilk}^t$ is a constant vector chosen so that

$$\underline{\alpha}_{ilk}^t \underline{\hat{\psi}} = \mu + \alpha_i + \hat{\beta}_l + \gamma_k$$

The variance of $\hat{\theta}_{ilk}$ can be approximated using the delta-method as

$$var(\theta_{ilk}) \approx \left(\frac{\partial \hat{\theta}_{ilk}}{\partial \hat{\eta}_{ilk}}\right)^2 var(\ln \theta_{ilk}) = \hat{\theta}_{ilk}^2 var(\ln(\hat{\theta}_{ilk})) \tag{4.3}$$

where $\hat{\eta}_{ilk} = \underline{\alpha}_{ilk}^t \underline{\hat{\psi}} = \ln \hat{\theta}_{ilk} - \ln N_{ilk}$ and

$$\begin{aligned}
& var(\ln(\hat{\theta}_{ilk})) \\
&= var(\underline{\alpha}_{ilk}^t \underline{\hat{\psi}}) \\
&= \underline{\alpha}_{ilk}^T cov(\underline{\hat{\psi}}) \underline{\alpha}_{ilk} \\
&= var(\hat{\mu}) + var(\alpha_i) + var(\hat{\beta}_l) + var(\gamma_k) \\
&\quad + 2cov(\mu, \alpha_i) + 2cov(\mu, \beta_l) + 2cov(\alpha_i, \beta_l) \\
&\quad + 2cov(\hat{\mu}, \gamma_k) + 2cov(\alpha_i, \gamma_k) + 2cov(\hat{\beta}_l, \gamma_k)
\end{aligned}$$

The variance of the predicted number of deaths at time l is

$$var(\hat{\theta}_l) = \sum_i \sum_k var(\hat{\theta}_{ilk}) + \sum_{(i,k)\neq(i',k')} cov(\theta_{ilk}, \hat{\theta}_{i'l'l}) \tag{4.4}$$

$$cov(\hat{\theta}_{ilk}, \hat{\theta}_{i'lk'})$$

$$\approx \frac{\partial \hat{\theta}_{ilk}}{\partial \hat{\eta}_{ilk}} \frac{\partial \hat{\theta}_{i'lk'}}{\partial \hat{\eta}_{i'lk'}} cov(\underline{\alpha}_{ilk}\hat{\underline{\psi}}, \underline{\alpha}_{i'lk'}\hat{\underline{\psi}})$$

$$= \hat{\theta}_{ilk}\hat{\theta}_{i'lk'} \left[cov(\hat{\mu} + \hat{\alpha}_i + \hat{\beta}_l + \hat{\gamma}_k, \hat{\mu} + \hat{\alpha}_{i'} + \hat{\beta}_l + \hat{\gamma}_{k'}) \right]$$

$$= \hat{\theta}_{ilk}\hat{\theta}_{i'lk'} \left[var(\hat{\mu}) + cov(\hat{\mu}, \hat{\alpha}_{i'}) + cov(\hat{\mu}, \hat{\beta}_l) + cov(\hat{\mu}, \hat{\gamma}_{k'}) + cov(\hat{\mu}, \hat{\alpha}_i) \right.$$

$$+ cov(\hat{\alpha}_i, \hat{\alpha}_{i'}) + cov(\hat{\alpha}_i, \hat{\beta}_l) + cov(\hat{\alpha}_i, \hat{\gamma}_k) + cov(\hat{\mu}, \hat{\gamma}_k) + cov(\hat{\alpha}'_i, \hat{\gamma}_k)$$

$$+ cov(\gamma_k, \hat{\gamma}_{k'}) + cov(\hat{\gamma}_k, \beta_l) + cov(\hat{\mu}, \hat{\beta}_l)$$

$$\left. + cov(\hat{\alpha}_{i'}, \hat{\beta}_l) + cov(\hat{\alpha}_{i'}, \hat{\beta}_l) + var(\hat{\beta}_l) \right] \tag{4.5}$$

The asymptotic 95% prediction intervals will be based on the standard errors using an assumption of normality (estimate $\pm(1.96\times$ standard error of the estimate)). The estimated SE are obtained on substituting the MLE of Ψ in (4.5).

5. Model Selection

Clytton and Schifflers (1987a and 1987b) advised the use of a reduced age-period (AP) or age-cohort (AC) model whenever possible, and the use of the full age-period-cohort (APC) model only when neither of these provides a satisfactory fit. In cancer epidemiology, however, this is often the case (Clytton and Schifflers 1987a; Clytton and Schifflers 1987b; Post *et al.* 1999). Since age is the most important predictor of prostate cancer mortality (Morrison *et al.* 1995), only models that include age will be considered in selecting a model to summarize the observed cancer rates. The two possible alternatives for model selection are an age-period-cohort model or an age-period model (AP). The AP model has the form

$$\log(E(n_{ij})) = \mu + \alpha_i + \beta_j + \log(N_{ij}),$$

where α_i are the age effects, β_j are period effects, $\log(N_{ij})$ is an offset and $E(n_{ij})$ expected number of mortality cases.

One way to confirm the effectiveness of a given model is to establish how it would have performed in predicting data that we have already observed. To evaluate the behavior of the APC models, we have examined prostate cancer mortality data from 1950 to 1979 as a means of predicting the mortality rates for the period of 1980 to 1990. The choice of the form of the model could have a profound effect on the forecasts. Accordingly, we evaluated two possible forms, with the aims of parsimony, scientific validity, and a reasonable fit to existing data.
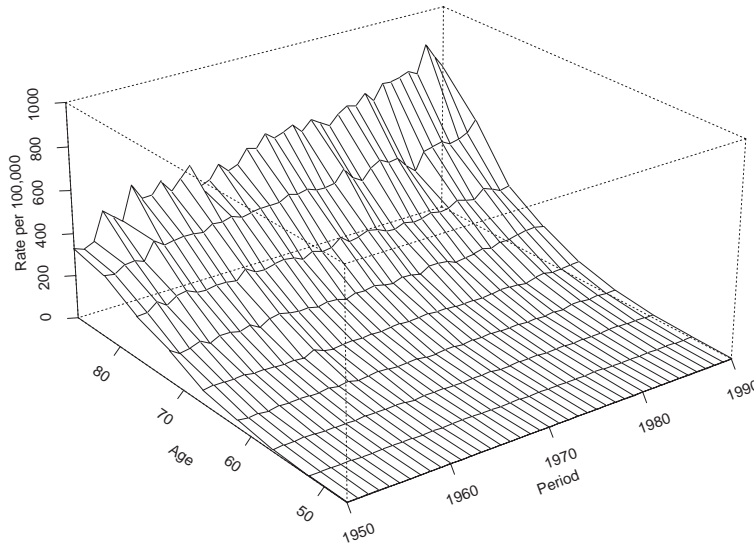
Figure 1: Canadian age-period specific mortality rates for prostate cancer, 1950-1990.
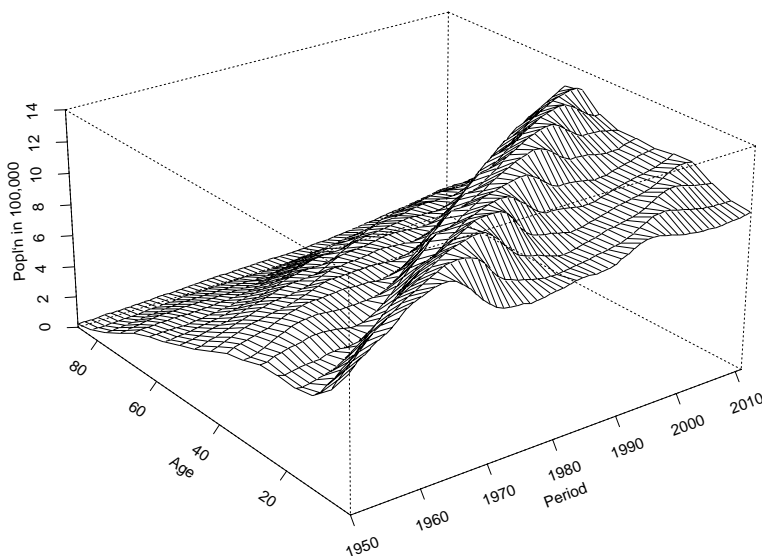

Figure 2: Canadian male age-period specific population, 1950-1990, with projections to the year 2010.

We chose two Poisson regression models: the AP and the APC models. The age-cohort model was not incorporated because we are interested in knowing the number of cases in the future per period but not per cohort, as it is unrealistic to extrapolate based on the few cases that appear in the most recent cohorts. Three indices of model performance were calculated:

1. Mean prediction error = mean (predicted number of cases − actual number of cases). This estimates the bias of the prediction. If negative, then the prediction is considered to be too low.

2. Mean absolute prediction error = mean (absolute (predicted number of cases − actual number of cases)). This index was used to provide the average difference from the actual observed cases.

3. Mean squared prediction error = mean ((predicted number of cases − actual number of cases)^2). This index of the performance incorporates both bias and variance. In this paper it is regarded to be more sensitive to outliers than the mean absolute error.

6. Example

Cancer is a leading health problem in Canada. Of all cancer among males, cancer of the prostate is the most common cancer diagnosed and is the third leading cause of death next to lung and colorectal cancers (National Cancer Institute of Canada 1993; Morrison et al. 1995). Because the Canadian population is growing and aging, the number of prostate cancer deaths is expected to continue to increase progressively (Figure 1). The age-period specific population data by five-year age groups for Canadian males for the time period 1950 to 1990 and extrapolated values to the year 2010 were obtained from Statistics Canada (Demography Division, 1993). It is worth mentioning that there is a large wave created by the baby boom of the period 1945-1970 (Figure 2) which is expected to influence health care delivery in the future as with regard to prostate cancer. We calculated the annual Canadian age-specific prostate cancer mortality rates among males aged 45 years or older for the period between 1950 and 1990 using 5-year intervals (i.e. 45-49 years to 85+ years). Age-specific rates were not calculated for men under the age of 45 because prostate cancer is very rare in this group (Figure 1). One-year periods were used to allow forecast for individual years. Cohorts were taken as the midpoints of five-year cohort intervals. This resulted in having 9 age groups, 40 periods and 17 cohorts.

7. Results

The two Poisson regression models fit separate parameters for each different age group and each year. The analysis of deviance shows evidence that all the effects (age, period and cohort) are nonzero ($p < 0.0001$, Table 1). Table 1 represents the effect of sequential inclusion each of the terms in the model, starting from the null model.

Table 1: Analysis of deviance table for age-period-cohort (APC) model

	Df[1]	Deviance[2]	Residual df	Residual deviance	Pr(Chi)[3]
Null[4]			368	153486.2	
Age	8	152239.0	360	1247.2	< 0.0001
Period	40	672.8	320	574.4	< 0.0001
Cohort	15	190.9	305	383.5	< 0.0001

[1]df = degrees of freedom

[2]Deviance: difference in residual deviance between each model and the one above it.

[3]Pr(Chi) = the tail probability (p-value) of the Chi-Squared distribution corresponding to the values in the "df" and "Deviance" columns. Small p-value indicates that there is not much evidence in favor of the null hypothesis, that the smaller model is correct, and it should be rejected.

[4]The null model is the mean of the response if an intercept (μ) is present in the model.

To forecast, both the period and cohort parameters were projected. Since the choice of periods is important for accuracy of the period parameters, we based the forecasted years on the most recent decade and on the 7 most recent cohorts only. Forecasts for the next ten years required an addition of ten period (1980, ..., 1990) and four cohort values (1933, 1938, 1943, and 1948). The new cohort values were not considered to be so crucial because it would be a long time before rates become large for young cohorts (Osmond and Gardner 1983). Based on the fitted 1950-1979 prostate cancer mortality data, the period effects and the cohort effects were both extended linearly to obtain the next ten period values and the next four cohort values. Recombining these values with the estimated age-values produced the projected rates, that is the projected rates, $\hat{\lambda}_{ilk}$, can be estimated by combining $\hat{\alpha}_i$, $\hat{\beta}_l$ and $\hat{\gamma}_k$ as $\hat{\lambda}_{ilk} = \hat{\mu} + \hat{\alpha}_l + \hat{\gamma}_k$.

Based on the three indices of model projections performance, the AP model seems to be able to make slightly better predictions than the APC model. However, as a result of the change of 1980-1990 mortality trends, both models produced very low predictions (Table 2).

Table 2: Results of the three indices of model projection performance

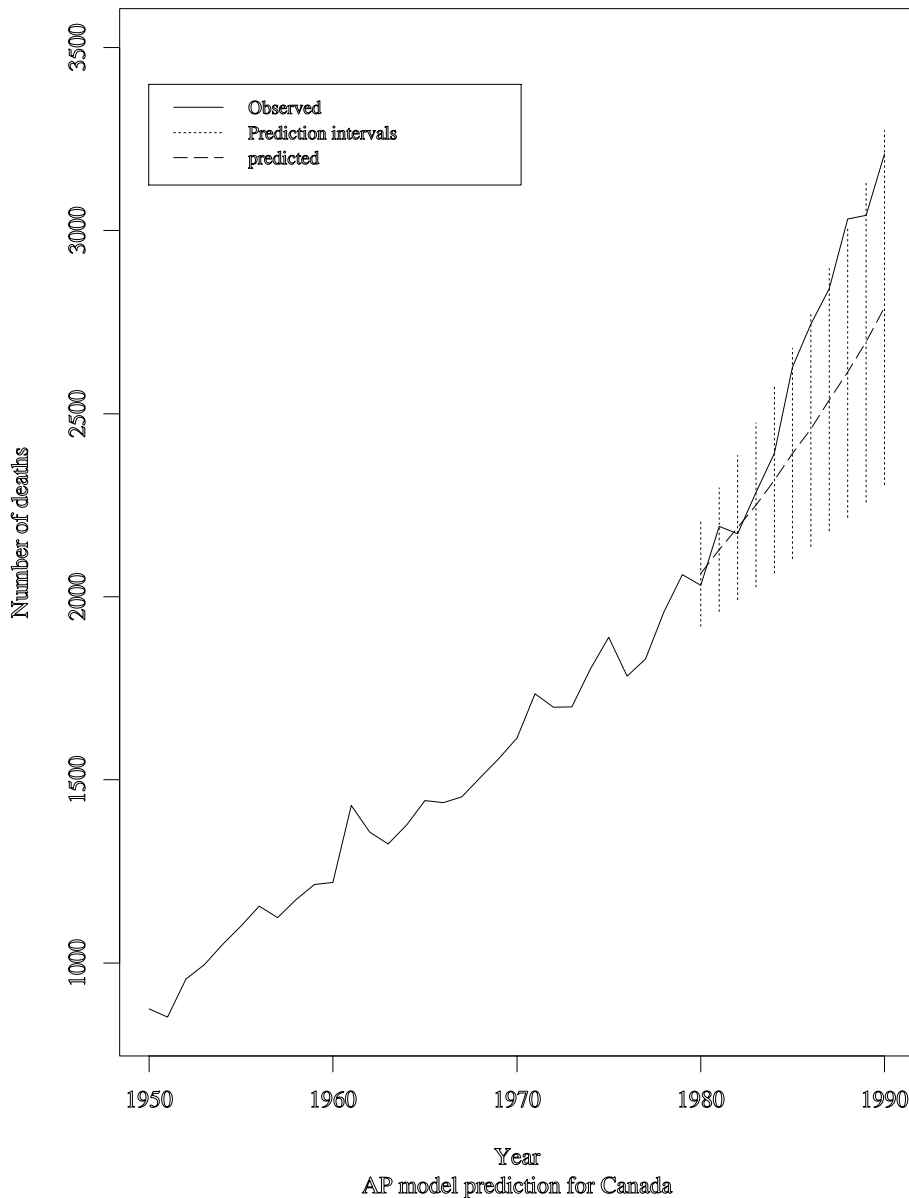| Model	Mean Error	Mean Absolute Error	Mean Squared Error
Age-period	−10.94	23.45	1350
Age-period-cohort	−18.98	25.49	1503

Figure 3: Observed and predicted number of deaths from prostate cancer in Canada based on the AP model. The vertical broken lines represent 95 percent prediction intervals for future observation accommodating the over-dispersion.

Table 3 and Figure 3 contain a summary of the results of fitting the AP model using 1950-1979 prostate data and predicting for 1980-1990. The AP model under-predicts the number of cases. The predicted values were within $\pm$ 3% of the observed values for the years 1980-1984. However, the AP model

under-predicted the number of cases of prostate cancer by around 10%.

Table 3: Observed and predicted numbers of prostate cancer deaths for 1980 to 1990.

Year	Observed cases	Predicted cases	95% prediction intervl
1980	2032	2064	1920 - 2208
1981	2192	2129	1960 - 2298
1982	2172	2189	1993 - 2385
1983	2285	2251	2027 - 2475
1984	2392	2319	2065 - 2573
1985	2627	2391	2104 - 2678
1986	2744	2458	2137 - 2779
1987	2842	2538	2180 - 2896
1988	3032	2614	2217 - 3011
1989	3042	2696	2258 - 3134
1990	3209	2790	2307 - 3273

8. Discussion

In this study, we fitted AP and APC models to prostate cancer mortality rates in Canada. Despite its nonidentifiability problem, the APC model appears suitable for forecasting the mortality due to prostate cancer. It has been shown that projections based on APC models can be uniquely determined and are not affected by the identifiability problem (Holford 1985). Regardless of the limitations of the APC model, our results show that it is useful in predicting the underlying trend of mortality rates due to prostate cancer in Canada.

Based on the three performance indices for model projections, the AP model appears to be able to slightly more successful in its predictions than the APC model. However, due to the changes in the mortality trends for 1980-1990, both models under-estimated the number of cases. This under-prediction of prostate cancer mortality may also be attributable to the observed baby boom of 1945-1970.

We have clearly demonstrated that accurate prediction intervals for the numbers of deaths from prostate cancer can be constructed using the AP model. A previous study (Hakulinen and Dyba 1994) discussed the prediction intervals for linear models assuming the Poisson distribution. Extra-Poisson variation is a particular problem in large geographical areas or with a common disease with a large number of cases (McCullagh and Nelder 1989). It will not make much difference to the estimates of the coefficients but it can have a considerable influence on the prediction intervals. Therefore, the prediction interval was adjusted

using a dispersion factor of 1.63 (McCullagh and Nelder 1989).

We conclude that the assumption of the Poisson distribution is not appropriate in describing the mortality of Canadian prostate cancer. The analysis of the deviance suggested over-dispersion and therefore a lack of fit, since the residual deviance was greater than the degrees of freedom. The assumption of a Poisson distribution for the number of cancer cases turns out to be unrealistic when Canada as a whole is considered. However, breaking down the data by small regions and/or socio-economic factors may improve the validity of these assumptions. Nevertheless, our study is one of only a few which have examined the values of forecasting by Poisson APC models. It differs from other studies in that it has identified the prediction intervals for new cases, addressed the practical problems of APC forecasts, and estimated the variance for the predicted number of cases using Poisson-distribution observations.

Acknowledgements

References

Bray, I. (2002). Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Applied Statistic* **51**, 151-164.

Clayton, D,, and Schifflers, E. (1987a). Models for temporal variation in cancer rates I: age-period and age-cohort models. *Statistics in Medicine* **6**, 449-467.

Clayton, D., and Schifflers, E. (1987b). Models for temporal variation in cancer rates II: age-period-cohort models. *Statistics in Medicine* **6**, 469-481.

Dyba, T. and Hakulinen, T. (2000). Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Statistics in Medicine* **19**, 1741-1752.

Hakulinen, T. and Dyba, T. (1994). Precision of incidence predictions based on Poisson distributed observations. *Statistics in Medicine* **13**, 1513-1523.

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

Holford, T. R. (1985). An alternative approach to statistical age-period-cohort analysis. *Journal of Clinical Epidemiology* **38**, 831-836.

Holford, T. R. (1991). Understanding the effects of age, period and cohort on incidence and mortality rates. *Annual Reviews of Public Health* **12**, 425-457.

Jee, S. H., Kim, I. S., Suh, I., Shin, D. and Appel LJ. (1998). Projected mortality from lung cancer in South Korea, 1980-2004. *International Journal of Epidemiology* **27**, 365-369.

Kupper, L. L., Janis, J. M., Salama, I. A., Yoshizawa, C. N. and Greenberg, B. G. (1983). Age-period-cohort analysis: An illustration of the problems in assessing interaction in one observation per cell data. *Communication in Statistics, Theory and Methods* **12**, 2799-2807.

Kupper, L. L., Janis, J. M., Karmous, A. and Greenberg, B. G. (1985). Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Diseases* **38**, 811-830.

MacNeill, I. B., Elwood, J. M., Miller, D. and Mao, Y. (1995). Trends in mortality from melanoma in Canada and prediction of future rates. *Statistics in Medicine* **14**, 821-839.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and Hall.

Morrison, H. I., MacNeill, I. B., Miller, D., Levy, I,, Xie. L. and Mao, Y. (1995). The Impending Canadian prostate cancer epidemic. *Canadian Journal of Public Health* **86**, 274-278.

National Cancer Institute of Canada (1994). *Canadian Cancer Statistics 1993.* Toronto.

Negri, E., La Vecchia, C., Levi, F., Randriamiharisoa, A., Decarli, A. and Boyle, P. (1990). The application of age, period and cohort models to predict Swiss cancer mortality. *Journal of Cancer Research and Clinical Oncology* **116**, 207-214.

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

Osmond, C. and Gardner, M. J. (1983). Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine* **1**, 245-259.

Post, P., Straatman, H., Kiemeney, L. and Coebergh, J. W. (1999). Increased risk of fatal prostate cancer may explain the rise in mortality in the Netherlands. *International Journal of Epidemiology* **28**, 403-408.

Schaubel, D., Morrison, H., Desmeules, M., Parsons, S. and Fenton, S. (1998). End-stage renal disease projections for Canada to 2005 using Poisson and Markov models. *Internation Journal of Epidemiology* **27**, 274-281.

Naser Elkum
King Faisal Specialist Hospital and Research Centre
MBC #3, P.O. Box 3354
Riyadh 11211
Kingdom of Saudi Arabia
e-mail: nkum@kfshrc.edu.sa