Bayesian Projections: What Are the Effects of Excluding Data from Younger Age Groups?

A. Baker¹ and I. Bray²

¹ School of Mathematics and Statistics, University of Plymouth, Plymouth, United Kingdom.
² Department of Social Medicine, University of Bristol, Bristol, United Kingdom.

Received for publication September 23, 2004; accepted for publication May 18, 2005.

Bayesian age-period-cohort models are used increasingly to project cancer incidence and mortality rates. Data for younger age groups for which rates are low are often discarded from the analysis. The authors explored the effect of excluding these data, in terms of the precision and accuracy of projections, for selected cancer mortality data sets. Projections were made by using a generalized Bayesian age-period-cohort model. Smoothing was applied to each time scale to reduce random variation between adjacent parameter estimates. The sum of squared standardized residuals was used to assess the accuracy of projections, and 90% credible intervals were calculated to assess precision. For the data sets considered, inclusion of all age groups in the analysis provided more precise age-standardized and age-specific projections as well as more accurate age-specific projections for younger age groups. An overall improvement in the accuracy of age-standardized rates was demonstrated for males but not females, which may suggest that analysis of the full data set is beneficial when projecting cancer rates with strong cohort effects.

Bayesian analysis; mortality; neoplasms; prediction; projection

Abbreviations: APC, age-period-cohort; SSRs, squared standardized residuals.

Much work has been undertaken in recent decades with the aim of producing projections of future cancer incidence and mortality rates from observed rates by using age-period-cohort (APC) models. This area of study is of particular importance with regard to estimating current cancer rates, because of the delay in compiling and publishing these data, and to public health planning for the future. Although age-standardized projections are a convenient summary measure of cancer incidence or mortality, facilitating comparison between countries and over time, they may mask important differences between age groups (1). Age-specific projections are therefore also of interest to epidemiologists and those planning future health services (2).

Classical APC models have been widely used to make cancer projections (3–5). In recent years, a Bayesian APC approach to projecting cancer rates has been implemented, which incorporates prior information about smoothness on each time scale to reduce random variation and improve the precision of the projections. These models are described in detail elsewhere (6–8). Cancer incidence and mortality data are typically given as counts stratified into 5-year age groups and 5-year calendar periods. There is little published evidence to suggest from which age group the analysis should begin (9). This paper examines the issue of discarding data from young age groups from a Bayesian APC analysis and the effect it has on the resulting projections.

In classical analyses, zero counts in the observed data may lead to problems in model fitting, such as unstable parameter estimates. This problem has been discussed in terms of generalized linear models (10) and specifically...
APC models (6). In some analyses, it has been assumed that younger age groups in which very few events are observed can be excluded from Bayesian APC analyses with little effect on the projections of interest (9, 11, 12). However, since zero counts do not pose implementational problems when fitting APC models in the Bayesian framework, others (7, 8, 13–15) have included all data for the youngest age groups (e.g., age 0–4 years).

It has already been suggested that early observations may strongly influence the width of the credible intervals around the projections (6). It has also been reported that when primarily zero counts are given for younger age groups, then discarding these data leads to projections similar to those based on all age groups, but that differences in results are found when deaths are commonly observed among younger age groups (6). In an analysis of lung cancer mortality, projections for recent birth cohorts differed markedly depending on whether data for younger age groups were included (16). We postulate that, even when observed rates in the younger age groups are very low, where a strong cohort effect exists (such as is typically observed in smoking-related cancers), data from younger age groups will be important to improve the accuracy of projections. The purpose of this paper is to explore the effect on projections of excluding young age groups displaying negligible mortality rates, in terms of both accuracy and precision, in the application of Bayesian APC models.

MATERIALS AND METHODS

Contrary to the trend in North America and western Europe, smoking-related cancers are currently increasing in some countries in central Europe, with a particularly high smoking prevalence and tobacco-related mortality observed in Hungary (17). We considered three smoking-related cancer sites: larynx, esophagus, and oral cavity (including pharynx). Data for Hungary were extracted from the World Health Organization mortality database for 15 age groups (0–4, ..., 70–74 years) and for six calendar periods from 1965–1969 to 1990–1994. In a previous analysis based on age 30–74 years only (12), projections were made for a further three calendar periods (i.e., 15 years) up to 2005–2009. Here, we compare 15-year projections based on data for all ages (0–74 years) with those based on ages 30–74 years.

Projections were made by using a generalized Bayesian APC model (6–8), implemented in the BUGS software program (18). The model describes mortality (or incidence) rates in terms of additive effects for age, period, and cohort by using a log-link function. Period effects are those that apply to an entire population at a certain point in time, while cohort effects are those that affect a group of people born around the same time. The Bayesian formulation allows information to flow from observed into projected periods (11). The program and further statistical details are provided in the Appendix.

For our initial analysis, projections were made for each of the cancer sites by using all of the age groups available (0–4, ..., 70–74 years). Then, the analyses were repeated with the younger age groups (0–4, ..., 25–29 years) excluded. Projected mortality rates based on the full and reduced data sets were compared by means of age-specific plots. The fit of the projections to rates observed in 1995–1999 was summarized across all age groups by calculating the sum of squared standardized residuals (SSRs), and age-standardized rates were calculated by using the world standard population (19).
# TABLE 1. Posterior estimates of fitted age-standardized mortality rates and 15-year projections* per 100,000 for selected cancers in Hungary

<table>
<thead>
<tr>
<th>Calendar period</th>
<th>Reduced data set †</th>
<th>Full data set ‡</th>
<th>Calendar period</th>
<th>Reduced data set †</th>
<th>Full data set ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Esophagus</td>
<td>Larynx</td>
<td>Oral cavity</td>
<td>Esophagus</td>
<td>Larynx</td>
</tr>
<tr>
<td>Males</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1965–1969</td>
<td>1.98</td>
<td>3.73</td>
<td>2.72</td>
<td>1.98</td>
<td>3.74</td>
</tr>
<tr>
<td>1970–1974</td>
<td>2.16</td>
<td>4.49</td>
<td>3.57</td>
<td>2.16</td>
<td>4.52</td>
</tr>
<tr>
<td>1975–1979</td>
<td>2.84</td>
<td>5.14</td>
<td>5.04</td>
<td>2.84</td>
<td>5.16</td>
</tr>
<tr>
<td>1980–1984</td>
<td>4.23</td>
<td>5.71</td>
<td>8.10</td>
<td>4.23</td>
<td>5.73</td>
</tr>
<tr>
<td>1985–1989</td>
<td>5.95</td>
<td>7.00</td>
<td>11.47</td>
<td>5.96</td>
<td>7.01</td>
</tr>
<tr>
<td>2000–2004</td>
<td>13.71 (8.24, 22.14)</td>
<td>14.07 (9.81, 19.87)</td>
<td>35.10 (22.05, 56.63)</td>
<td>13.75 (8.35, 22.03)</td>
<td>13.74 (9.62, 19.41)</td>
</tr>
<tr>
<td>2005–2009</td>
<td>18.08 (6.26, 49.82)</td>
<td>17.97 (8.44, 37.49)</td>
<td>50.86 (19.63, 136.71)</td>
<td>18.10 (6.38, 49.26)</td>
<td>16.87 (7.90, 34.95)</td>
</tr>
<tr>
<td>Females</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1965–1969</td>
<td>0.37</td>
<td>0.33</td>
<td>0.52</td>
<td>0.37</td>
<td>0.33</td>
</tr>
<tr>
<td>1970–1974</td>
<td>0.36</td>
<td>0.33</td>
<td>0.58</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td>1975–1979</td>
<td>0.37</td>
<td>0.36</td>
<td>0.70</td>
<td>0.38</td>
<td>0.36</td>
</tr>
<tr>
<td>1980–1984</td>
<td>0.42</td>
<td>0.41</td>
<td>0.91</td>
<td>0.43</td>
<td>0.42</td>
</tr>
<tr>
<td>1985–1989</td>
<td>0.52</td>
<td>0.50</td>
<td>1.25</td>
<td>0.52</td>
<td>0.51</td>
</tr>
<tr>
<td>1990–1994</td>
<td>0.67</td>
<td>0.61</td>
<td>1.82</td>
<td>0.67</td>
<td>0.63</td>
</tr>
<tr>
<td>1995–1999</td>
<td>0.90 (0.75, 1.10)</td>
<td>0.77 (0.64, 0.93)</td>
<td>2.67 (2.33, 3.08)</td>
<td>0.88 (0.73, 1.08)</td>
<td>0.79 (0.65, 0.94)</td>
</tr>
<tr>
<td>2000–2004</td>
<td>1.23 (0.85, 1.81)</td>
<td>0.98 (0.69, 1.40)</td>
<td>3.86 (2.90, 5.13)</td>
<td>1.15 (0.81, 1.71)</td>
<td>0.98 (0.69, 1.41)</td>
</tr>
<tr>
<td>2005–2009</td>
<td>1.68 (0.82, 3.61)</td>
<td>1.28 (0.64, 2.60)</td>
<td>5.42 (3.05, 9.60)</td>
<td>1.49 (0.74, 3.23)</td>
<td>1.25 (0.64, 2.46)</td>
</tr>
</tbody>
</table>

* Mortality rates for the last three calendar periods, followed by 90 percent credible intervals in parentheses.
† Data for persons aged 0–29 years were not included in the analysis.
‡ The analysis included data for persons aged 0–74 years.
RESULTS

Both age-specific and age-standardized projected rates were compared by using the full and the reduced data sets. An example of an age-specific plot (larynx cancer mortality for males aged 30–34 years) is shown in Figure 1. The most striking feature is the increased width of the 90 percent credible intervals based on the reduced compared with the full data set, particularly for longer term (i.e., 15-year) projections. Plots for the other data sets considered were similar, and, in each case, age-specific projections for ages 30–34 years were more accurate when based on the full data set. In progressively older age groups, the differences became less apparent, and the fitted and projected rates for the two sets of results tended toward each other.

From the age-standardized mortality rates given in Table 1 we see that the fitted and projected rates were similar when the full and the reduced data sets were used. Although there was no systematic difference in estimates based on the two models for 5- and 10-year projections, 15-year projections tended to be lower when the full data sets were analyzed. As expected, the credible intervals for the three projected periods were generally narrower when all of the age groups were included in the analyses, and differences in the width of the credible interval became greater the longer term the projections were. Thus, there may be little or no difference in credible interval width when projecting one 5-year period; however, when projecting 15 years into the future, the credible intervals based on all age groups are distinctly narrower.

The accuracy of projections for the subsequently observed period 1995–1999 was assessed by summing SSRs across age groups (Table 2). These results confirmed that, as noted above, the fit of the 5-year projections was similar when the full and the reduced sets of data were used. Overall, the full data sets led to projections closer to the observed values for two of the three cancer sites (larynx and oral cavity). When the SSRs for males and females were summed across cancer sites, the full data sets produced the most accurate projections for males, while the opposite was true for females.

DISCUSSION

We compared projections of cancer mortality based on a Bayesian APC model with and without data from the youngest age groups included in the analysis. We used the sum of SSRs to assess the accuracy of 5-year projections and used 90 percent credible intervals to assess the precision of projections for up to 15 years based on data from a 30-year period. Our results suggest that including all age groups in an analysis reduces the width of credible intervals for age-specific and age-standardized rates (i.e., produces projections of greater precision) and that this benefit increases with length of projection. On the basis of our findings, age-specific projections based on full data sets are more accurate for the younger age groups. Evidence for an improvement in accuracy is less consistent in age-standardized projections because the contribution to the sum of SSRs from the younger age groups is dominated by larger residuals in the older age groups. In the examples considered here, using the full data sets was found to improve accuracy of age-standardized projections for males but not for females. Our results suggest that including younger age groups will improve projections if the data contain information about important cohort effects, such as were observed in these data sets for males (12). In other cases, it could be argued that including childhood cases in cancer projection analyses would result in less accurate projections because of differences in childhood and adult etiology, for example, acute lymphoblastic leukemia (20).

There is still much work to be done in determining the best modeling approach in different situations. For example, further methodological work could develop optimal methods of making short- and long-term projections. While short-term projections are important for allocation of resources in the immediate future, long-term projections play a key role in estimating the burden of cancer in the coming decades (21) and prompting governments to take action to prevent predicted increases in morbidity and mortality.

We explored whether the magnitude of the excluded rates dictates whether their exclusion will affect projections, as suggested previously (6). We tabulated both the total number of cases excluded from the analyses and the rate for the last observed calendar period for the eldest excluded age group (these tables are available on request from the authors), but we were unable to identify any associations between these factors and our observations on the accuracy and precision of projections based on full and reduced data sets.

On the basis of our results and conclusions, we have several recommendations for further work. Firstly, when we compared results from the analysis of the reduced data sets for Hungary with those previously published (12), we noticed some small discrepancies in both fitted and projected rates. The maximum percentage difference observed...
was 3 percent (laryngeal cancer in females, 1965–1969); all other differences were of the order of 1 percent or less. This discrepancy is probably due to the process of sampling inherent in the methodology but should nevertheless be overcome by using sufficiently large numbers of iterations. This finding indicates that there is scope for improving convergence and mixing in applying these models.

Secondly, 15-year projections were consistently lower when the full data set was used (with the exception of cancer of the esophagus in males). Therefore, the central European data (12) should be reanalyzed, including all age groups and data for the period 1995–1999.

Finally, there is a similar ambiguity with regard to which older age groups should be included in the analysis. Since death certification in the very old is often unreliable (22), it is reasonable to exclude age groups above a certain point. However, as with the young age groups, the cutoff point varies between analyses. In the Bayesian literature, some authors have used an open category of 85 years or older (6, 7, 9, 16) as the highest age group. Others have used an upper limit of 74 years (8, 12, 14, 15) or 80 years (11). Since, in these models, the highest age groups feed information into neighboring age groups, the discarded data for older age groups may influence projected rates. This area warrants further investigation.

In conclusion, our results suggest that using all available data may improve Bayesian APC projections when there are strong cohort effects and when etiology is similar across all ages.

**ACKNOWLEDGMENTS**

Conflict of interest: none declared.

**REFERENCES**


**APPENDIX**

For each age group $a$ ($a = 1, 2, \ldots, A$) and period $p$ ($p = 1, 2, \ldots, P$), we have the number of deaths and the number of person-years ($n_{ap}$). We assume that the number of deaths is Poisson distributed ($\mu_{ap}$) and that $\log(\mu_{ap}) = \log(n_{ap}) + \alpha_a + \beta_p + \gamma_c$, where the cohort $c = 1, 2, \ldots, C$ is the approximate year of birth. Note that problems with identifying individual parameters in the model, caused by the linear association between age, period, and cohort effects, do not affect projected rates. To improve numeric stability and mixing, the linear trend of the age effects $\alpha_a$ is constrained to zero, and the period effects $\beta_p$ and cohort effects $\gamma_c$ are made to sum to zero by subtracting their means.

An autoregressive prior smoothing model is applied to each time scale to reduce variation between adjacent parameter estimates. We define prior distributions for age, period, and cohort effects that smooth each point on the two preceding points (random walk 2). The prior distribution for the age effects is
Prior distributions for period and cohort effects are similar. The hyperparameters $\sigma^2_\alpha$, $\sigma^2_\beta$, and $\sigma^2_\gamma$, which control the degree of smoothing on each time scale, are given noninformative $\text{gamma}(0.001,0.001)$ priors.

This model is implemented in BUGS, and the program is given below. Fitted and projected rates are estimated from samples of 10,000 drawn from the posterior distribution after excluding the first 1,000 iterations as “burn-in.” The samples are summarized by their median and 90 percent credible interval, calculated from the 5th and 95th percentiles.

The BUGS program follows:

```plaintext
model Ptwo;

### Specify constants. These can be changed depending on the data
### set and how many periods you wish to project ###
const
  I=90, A=15, P=6, C=20, PP=9, CC=23;

### List of variables in the model ###
var
  age[I], period[I], cohort[I], cases[I], pyr[I], mu[I],
  alpha[A], gamma[PP], beta[CC], alphac[A], gammac[PP], betac[CC],
  alphamean[A], gammamean[PP], betamean[CC], alphaprec[A],
  gammaprec[PP], betaprec[CC], taua, taup, tauc, ivec[A],
  aivec[A], rate[A,PP];

### Name data file and columns it contains ###
data cases, pyr, age, period, cohort in “larynxm.dat”;

### File of initial values ###
inits in “inits.in”;

### loop over all observations ###
{ for (i in 1:I) {

### define age-period-cohort model ###
cases[i] ~ dpois(mu[i]);
log(mu[i]) <- log(pyr[i]) + alphac[age[i]] + gammac[period[i]] +
  betac[cohort[i]];
}

### autoregressive prior model for age effects ###
alphamean[1] <- 0.0;
alphaprec[1] <- taua*1.0E-6;
alphamean[2] <- 0.0;
alphaprec[2] <- taua*1.0E-6;
for (a in 3:A) {
alphamean[a] <- 2*alpha[a-1] - alpha[a-2];
alphaprec[a] <- taua;
}
```
### loop over age groups ###
for (a in 1:A) {

### sample age effects ###
alpha[a] ~ dnorm(alphamean[a], alphaprec[a]);
}

### sample precision for smoothing of age effects ###
taua ~ dgamma(1.0E-3, 1.0E-3);

### remove linear trend from the age effects ###
for (a in 1:A) {
  ivec[a] <- a - (A+1)/2;
  aivec[a] <- ivec[a] * alpha[a];
  alphac[a] <- alpha[a] / (ivec[a] * sum(aivec[1]) / (A*(A+1)*(A-1)/12);
}

### autoregressive prior model for period effects ###
gammamean[1] <- 0.0;
gammamean[2] <- 0.0;
gammamean[p] <- taup;
for (p in 3:PP){
  gammamean[p] <- taup;
}

### loop over periods ###
for (p in 1:PP){

### sample period effects and subtract mean for observed periods #
gamma[p] ~ dnorm(gammamean[p], gammaprec[p]);
gammac[p] <- gamma[p] - mean(gamma[1:P]);
}

### sample precision for smoothing of period effects ###
taup ~ dgamma(1.0E-3, 1.0E-3);

### autoregressive prior model for cohort effects ###
betamean[1] <- 0.0;
betamean[2] <- 0.0;
betamean[c] <- 2*beta[c-1] - beta[c-2];
betamean[c] <- tauc;
for (c in 3:CC){
  betamean[c] <- 2*beta[c-1] - beta[c-2];
  betamean[c] <- tauc;
}

### loop over cohorts ###
for (c in 1:CC){
### sample cohort effects and subtract mean for observed cohorts #
beta[c]~dnorm(betamean[c],betaprec[c]);
betac[c]<-beta[c]-mean(beta[1:C]);

### sample precision for smoothing of cohort effects ###
tauc~dgamma(1.0E-3,1.0E-3);

### loop over age groups and periods ###
for(a in 1:A){
  for(p in 1:PP){

### calculate fitted and projected rates per 100,000 ###
rate[a,p]<-100000*exp(alphac[a]+gammac[p]+betac[A+p-a])
  }
}