A simple benchmark for mesothelioma projection for Britain

MARÍA DOLORES MARTÍNEZ MIRANDA†
BENT NIELSEN‡§
JENS PERCH NIELSEN¶

September 2015

Abstract

Background: It is of considerable interest to forecast the future burden of mesothelioma mortality. Data on deaths are available, whereas no measure of asbestos exposure is available.

Methods: We compare two Poisson models. First, a response-only model with an age-cohort specification. Second, a dose-response model using a synthetic exposure measure.

Results: The response-only model has 5% higher peak mortality than the dose-response model. The former performs better in out-of-sample comparison.

Conclusion: Mortality among males below 90 years of age is predicted to peak at 2079 deaths in 2017. The response-only model is a simple benchmark that forecasts just as well as more complicated models.

1 Background

The number of UK mesothelioma deaths continues to increase with 2049 male deaths below 90 years of age in 2013. It is of considerable interest to forecast the future burden of mesothelioma deaths. Mesothelioma is a cancer that is thought to be caused by exposure to asbestos fibres. It has a long latency period and mainly affects men. Once discovered it is rapidly fatal. The use of asbestos has been regulated in the UK since 1969, but substantial asbestos exposure continued until the mid 1980s. Based on a case-control study Rake et al. (2009) find that most metholioma cases are caused by occupational exposure. It is possible that non-occupational exposure to asbestos fibres may be increasing, see Carbone et al. (2012). In any case, while the UK has good records of mesothelioma deaths there is no data on the exposure to asbestos fibres.

For the UK a number of forecasts of mesothelioma mortality are available. The lack of exposure data has been dealt with in different ways. Hodgson et al. (2005) used a dose-response model. They pioneered a method that constructed a synthetic exposure measure based on epidemiological clearance model involving the half-time of clearing of asbestos fibres from the lungs along with measures of asbestos import to the UK. Tan et al. (2010) and the Health Service Executive (2014) updated the analysis using a Bayesian set-up, see also Tan and Warren (2009). Martínez Miranda et al. (2015) suggested a
response-only method. This is a simpler method, requiring less epidemiological insight. Hence, it is less dependent on uncertainty in this insight. The latter projection gives a somewhat higher estimate for the future burden. Here we compare the methods in the light of the most recent data.

2 Materials and methods

2.1 The data

The Health Service Executive routinely publishes data with 5-year age groups, most recently for 1968-2013, see Health Service Executive (2015). The Health Service Executive has kindly provided us with a break down by annual age groups. Most deaths are male (85%) and in the age groups 25-89 (99%). In recent years there has been an increasing number of deaths at age 90+. For instance for 2013 there are 74 such deaths. To be consistent with the literature we use ages 25-89 for men.
Figure 2.1 shows summary plots of the number of mesothelioma deaths. The first three plots show observed counts of deaths by age, period and cohort. In particular, the second plot shows how mortality continues to rise. The fourth plot is a sparsity plot indicating age-period combinations with only few incidences of death. This illustrates that most mesothelioma cases are among those of age above 45. Indeed, there are only 45 cases for the 1967-1988 cohorts, which have benefited most from predictive measures since the 1970s. It can also be discerned that the number of cases is increasing among those of age above 70.

In summary, the data set is an age-period array of cases of mesothelioma deaths. No data is available on individuals’ exposure to asbestos and hence the person years at risk. We have annual information for ages 25–89 and periods 1968–2013, which is used for estimation and forecasting.

### 2.2 The model without exposure

Martínez Miranda et al. (2015) proposed a model that does not use exposure. It is a Poisson regression model for the counts of deaths with an age-cohort structure. This is inspired by the chain ladder analysis that is used for forecasting liability reserves in general insurance, see England and Verrall (2002). Inference is done by conditioning on the overall number of observed deaths. This appeals to a classic statistical result that conditioning in a Poisson distributed contingency table results in a multinomial sampling scheme.

In this model, the expected number of deaths is

$$EY_{\text{age,period}} = \exp(\alpha_{\text{age}} + \gamma_{\text{cohort}}),$$

(2.1)

where $Y_{\text{age,period}}$ is the number of deaths at a given age and a given period. We have $\text{age} + \text{cohort} = \text{period}$, so that $\alpha_{\text{age}}$ are age parameters and $\gamma_{\text{cohort}}$ are cohort parameters. The model could potentially be extended with a period factor, which would give an age-period-cohort model. Both the age-cohort model (2.1) and the age-period-cohort model are over-parametrised. We therefore use the canonical parametrisation suggested by Kuang et al. (2008a), see also Nielsen (2014a). This expresses the model for the log mortality in terms of interpretable, identifiable and freely varying parameters. Having done this standard generalized linear model techniques can be used. Details are given in the appendix.

The fit of the model can be tested formally by comparing the log likelihood of the model with that of a saturated model where the mean of the number of deaths, $Y_{\text{age,period}}$, is unrestricted. Twice this difference in log likelihoods is called the deviance. The presence of a period effect can be tested by comparing the deviances of models with and without period effects through a $\chi^2$ log likelihood ratio test. This is also called a deviance test.

The method for forecasting future mortality depends on the forecast region. In the context of an age-cohort model there are two possibilities: first, forecast the mortality of cohorts that are in the sample; second, forecast the mortality both of in-sample cohorts and of future cohorts. In the mesothelioma context the first choice seems most appropriate since youngest cohorts have benefitted from protective measures in the
workplace. With the first method point forecasts can be constructed directly from the estimated canonical parameters. Distribution forecasts can be calculated numerically using asymptotic methods. With the second method it is necessary to extrapolate cohort parameters, which has to be done with care, see Kuang et al. (2008b) for a detailed discussion.

The forecasts can be robustified by employing an intercept correction, see Hendry and Nielsen (2007). The idea is to line the forecast up with the most recent observation to take possible shifts at the forecast origin into account. For the present model it corresponds to multiplying the point forecasts with a factor given by the number of observations in the last period, divided by the prediction for the last period.

2.3 Models with synthetic exposure measures

An alternative model is to construct a synthetic measure for exposure. In the early work by Peto at al. (1995) general population figures were used. In the work by Hodgson et al. (2005) a more intricate method is used. It is based on a multinomial frequencies, see (A.2) and utilizes an epidemiological clearance model for these frequencies. This model essentially replaces the log-linear age-cohort model with a non-linear model involving fewer parameters representing age-specific exposure, overall population exposure and a diagnostic trend measuring recording error, along with guesses of a background rate and of the half-life for asbestos clearance.

The model is estimated by Bayesian methods developed by Tan et al. (2010) for the Health Service Executive using data until 2006. The Health Service Executive (2014) provides updated forecasts using data until 2010. The Bayesian model results in distribution forecasts, where the point forecasts are somewhat lower than those of the response-only model, but with forecast intervals that are of the same order of magnitude.

Tan and Warren (2011) suggested another model based on a two-stage clonal expansion model, see also Asbestos Working Party (2011). While this gives somewhat lower forecasts, it gives alternative epidemiological insight in the future development of the epidemic.

3 Results

The annual number of mesothelioma deaths continues to rise and the peak has not been reached yet. We first compare the fits and then the forecasts of the two models without exposure and with synthetic exposure.

3.1 Fit

We fitted the model without exposure to the updated data 1968-2013. Table 3.1 shows deviances for the age-period-cohort model and the age-cohort model. It also shows the relative deviances of those models. Based on this evidence we maintain the age-cohort model. Likewise we do not find evidence against the model when considering plots of the residuals. Those plots are not reported here.
Table 3.1 also shows the deviance of the two-stage clonal expansion model. This is computed from the in-sample predictions reported by the UK Asbestos Working Party (2011) using the data for 1968-2008. Measured in terms of fit, the two-stage clonal expansion models appears to be just as good as the age-cohort model. The point is that in-sample fit may not be a guide to the quality of out of sample forecasts.

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Deviance</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1968-2013</td>
<td>APC</td>
<td>2763.6</td>
<td>2772</td>
<td>0.542</td>
</tr>
<tr>
<td></td>
<td>AC</td>
<td>2818.1</td>
<td>2816</td>
<td>0.485</td>
</tr>
<tr>
<td></td>
<td>AC vs. APC</td>
<td>54.6</td>
<td>44</td>
<td>0.132</td>
</tr>
<tr>
<td>1968-2008</td>
<td>Exposure</td>
<td>2651.9</td>
<td>2646</td>
<td>0.704</td>
</tr>
</tbody>
</table>

Table 3.1: Deviance analysis of the age-period-cohort model.

### 3.2 Forecast

Figure 3.1 shows sums of cases by period along with various forecasts. In the sums of cases we exclude cohorts from 1967 and later.
The observed number of cases are indicated with dots. We notice a large volatility in the most recent years, with a low 2011 count and a high 2012 count.

There are four forecasts. The forecast from the age-cohort model using data until 2013 is shown as a solid line. The shaded region indicates pointwise 95% forecast error bands. These are computed from asymptotic methods and include both process error and estimation error. The largest value of the point forecast is at 2079 in 2017 with 95% error band of (1998, 2161). The 2013 in-sample residual is very small, so intercept correction would increase the peak forecast modestly to 2083.

The dotted line is a forecast from the clearance models using data until 2010, reported by Health Service Executive (2015). This forecast includes cohorts until 1986. Nonetheless, they come out lower than the age-cohort forecasts. We do not show the forecast error bands as the width of the 90% error bands is nearly identical to width of the 95% bands of the age-cohort model reported in Figure 3.1. In particular, for 2013 the 90% error bands are (1901, 2069), as compared to the observed 2049 deaths.

The 2012 age-cohort model and the 2010 clearance model use different data, so they are not directly comparable. To illustrate their differences it is useful to consider forecasts from 2006: the dash-dot line is the age-cohort forecast with intercept correction, see Martínez Miranda et al. (2015), while the dashed line is the clearance model of Tan and Warren (2009). These forecasts generally have the same shape as their respective updates. The age-cohort forecast is most different, since the 2013 age-cohort forecast is affected by the recent volatility in the data. The in-sample quality of the two 2006 models are comparable, see Table 3.1. But out of sample the shape of the forecasts are different. The clearance model includes some choices of epidemiological tuning parameters. If these are chosen differently the clearance model forecasts will be more similar to the age-cohort forecasts, see Tan and Warren (2011). In conclusion, the recent data points appear to be more in line with the forecasts from the age-cohort model than those from the reported clearance model.

4 Conclusions

Dose-response models are difficult to use in mesothelioma projections due to the lack of data on exposure. A simple response model with an age-cohort structure avoids the need for constructing synthetic clearance based measures of exposure and it can be analyzed using standard generalized linear model techniques. From the UK data it appears that the response model gives forecasts that are as good as, if not better than, those from the more complicated clearance models. We therefore recommend using a simple response model as a benchmark for mesothelioma projections.

In the empirical analysis we find a peak of 2079 male deaths in 2017 for those of age less than 90 and cohort older than 1967. This is slightly worse and slightly later than the forecast of the clearance model.
The age-cohort model (2.1) has log mean $\mu_{age,period} = \alpha_{age} + \gamma_{cohort}$. Following Kuang et al. (2008a), Martínez Miranda et al. (2015) and Nielsen (2014a) we reparametrize it in terms of freely varying parameters involving an overall level of $\mu_{age,period}$ and differences or contrasts of the type $\alpha_{age} - \alpha_{age-1}$. We can then write $\mu_{age,period} = X'_{age,period}\xi$, where $X_{age,period}$ is a vector of the design matrix and $\xi$ is the freely varying canonical parameter. The log likelihood is therefore
\[
\log L(\xi|Y) = \sum_{age,period} \{Y_{age,period}\mu_{age,period} - \exp(\mu_{age,period})\}.
\] (A.1)

The model can be estimated by any standard generalized linear model routine, like the R package `apc`, see Nielsen (2014b). Martínez Miranda et al. (2015) develop pointwise forecast error bands using an asymptotic theory that takes into account both process error and estimation error. We used the R code from that paper.

The multinomial sampling scheme arises when conditioning on the total number of deaths. The multinomial frequencies are of the form
\[
\pi_{age,period} = \exp(\mu_{age,period})/\sum_{age',period'} \exp(\mu_{age',period'})
\] (A.2)

**Acknowledgements**

We thank A.J. Darnton at the Health Service Executive for providing the most recent version of the mesothelioma register. All numerical computations were done with R, see R Development Core Team (2014). M.D. Martínez Miranda is supported by project MTM2013-41383P from the Ministry of Economy and Competiveness and the European Regional Development Fund. B. Nielsen is supported by the Institute for Economic Modelling at the Oxford Martin School.

**References**


