Prediction of lung cancer rates 1986-2005

Comparison of several methods of prediction all based on O&G analysis, applied to males in 6 countries

Author : Mrs. B. Forey
Date : 8th February 1989

Introduction

This document provides a record of the work carried out during December 1988 and January 1989 to assess the prediction methods suggested by C.D. Gooch[1] and to explore variations on the methods. A brief description of this work, together with a full explanation of the underlying ideas and of the method eventually chosen, are given in the report on predictions in 24 countries[2], and most readers should refer to that report.

1. Previous work

Previous work by CDG[1] on lung cancer in England & Wales had suggested 4 methods of prediction based on O&G analysis. These were:

i) A-P-C model with linear extrapolation of P and C values

ii) A-P-C model with log-linear extrapolation

iii) A-C model with log-linear extrapolation of C values

iv) A-C model with log-linear extrapolation and residual correction.

Extrapolation of cohort values used a weighting of 0 for values before the peak and 1 for the peak and after. Extrapolation of the period values from methods i) and ii) and of residuals from iv) used weights decreasing exponentially into the past, to allow more recent values to have more influence than those in the distant past.

From the then available 1941-80 data, CDG used various (4, 5 or 6) initial periods to predict the remaining periods. He compared age-specific numbers of deaths. He found that predictions improved if more points were used, except that pre-1950 data did not improve predictions. The A-P-C model produced better results for males, but the A-C was better for females. Log-linear extrapolation gave poorer results than linear, but linear gave the possibility of negative results.

CDG then used the 4 methods to predict rates and numbers of deaths over the 20-year period 1981-2000.

2. CDG methods applied to 6 countries

As a quick "first look" method ii was applied to male lung cancer rates in the 24 countries we have studied. Only rates were studied, to avoid the need for population predictions. Since different lower age limits were used in the various countries, we concentrated on ages 40-74 since they are available throughout (except in four countries for females). The predicted rates, standardised to the England & Wales 1981-85 population, were presented at the meeting with BAT on 15/12/88, as a table, plot and rank diagram. It was emphasised that these were only preliminary results and that the stability of predictions needed looking at. In particular the method of weighting values after the cohort peak could only be applied in a pretty arbitrary manner in some countries - where there might be double peaks, a very recent peak, a plateau, or no peak at all. Six countries were selected to give a variety of patterns and possible problems, and the four methods suggested by CDG were applied. The countries are: Canada, Denmark, England & Wales, W. Germany, Greece, USA.

Plots comparing the 6 countries were done for methods ii) (Figure 1) and for iii) & iv) (Figure 2). These show observed rates for 1951-85 and predicted for 1986-2005. (The plot for ii) is therefore a reduced version of the plot presented to BAT). However these plots did not seem particularly helpful and all subsequent plots have been specific to the individual countries.

Two types of plot are used.

(1) For a particular model both linear (dotted) and log-linear (dashed) extrapolations are shown. The cohort and period values are shown in the top part of the plot,

and at the bottom the age-specific rates are shown, solid representing observed and long dashes as fitted by the model. Extrapolations and the resulting predicted rates are shown as linear: dots, log-linear: dashes and residual correction: short dashes.

(2) The 40-74 standardised rates from all methods are brought together for comparison. The observed values are shown as black 0, and the A-C and A-P-C fitted models as red 1 and green 3 respectively. The green hardly shows at all since a good fit is assured by the constraints of the model, depending on which age groups were included in the analysis. For the predictions, the same colours are used with solid line for log-linear, dashes for linear and dots for residual correction. The other lines will be explained later. The solid green 4 line is method ii) as presented to BAT and is considered as the baseline for comparisons. (Linear predictions stop short if the estimated cohort values become negative).

Looking at the 40-74 std plots (Figures 3-8), there is not much difference between methods i) and ii) for most countries (green 3 and 4). This is because the extrapolated cohort values are not needed until quite late in the predictions. For instance, in England & Wales (or any other country where the analysis is based on age 25-74), we have estimates in the model for cohorts 1-16 and have extrapolated values for cohorts 17-20. The age group 40-74 in various periods actually comprises the following cohorts:

1981-85	10-13
1986-90	11-14
1991-95	12-15
1996-00	13-16
2001-05	14-17

So it is not until the 4th predicted period that the first extrapolated cohort value enters the estimation of this rate. Meanwhile the period values are relatively flat and hence there is little difference between their linear and log-linear extrapolations.

In all countries except Denmark, the A-C model (red 1) predicts higher rates than the A-P-C model. This seems to be due to misfit of the model, overestimating at the last period. The residual correction method (red 2) therefore looks quite good, bringing the predictions closer to those of the A-P-C model, except for England & Wales where they drop below. However, the age-specific predictions do not look sensible in some countries, particularly at the younger ages, seen for example W. Germany.

The extrapolated values and age-specific rates for each country are shown in Figures 9-14 for the A-P-C model and Figures 15-20 for the A-C model.

3. Exploration of some ideas and problems

Some further ideas have also been explored and are also included in Figures 3-8

(A) We are using age-standardised rates for 40-74 as the best possible indicator for overall trends. For countries where there are sufficient numbers of deaths to include down to age 25, the predictions for age 40-74 depend chiefly on the assumption that cohort values estimated on the latest young people will continue to apply as they get older - the extrapolated values only come into use at the 4th period and only for one age-group (plus extrapolation of the period values which have a smaller effect but at all projected periods).

The assumption that cohort value, in some way related to cigarette consumption, may be determined at age 25-29 is questionable anyway for such a young age-group; see misfit for USA for example.

It seemed interesting, therefore, to use the alternative approach of getting cohort estimates based only on age 40-74, and to rely on extrapolations rather than estimates based on younger ages. This is effectively what has to be done in smaller countries anyway. This analysis has been done using A-P-C methods only - both linear and log-linear extrapolation, shown as blue plots 5 and 6. It has been done for England & Wales, W. Germany and USA, and similarly for Canada and Greece where original analyses were based on age 30+ (Figures 21-26).

The results of this are pretty much as expected for four of the countries - in England & Wales and USA, the predictions are higher since the youngest recent data (now excluded) had shown a particularly sharp decrease, in Canada there is little difference, and in Greece the predictions are lower. W. Germany: Excluding the younger data from the O&G analysis and thus excluding the latest cohorts removes the start of the cohort decline, and hence the extrapolation based on 40-74 analysis gives quite different cohort values than the 25-74 analysis. This leads to higher predictions which rise more steeply.

The 40-74 analysis approach was also used for Denmark, but here it illustrates a rather special factor: The original analysis was only based on age 35-74, so only one age group (and hence one cohort) have been excluded. However the final point of the full analysis was a particularly peculiar one, and due to the instability of the final estimate from O&G, we have in our discussions placed very little weight on it. But it heavily influences the extrapolation of the cohort values. The predictions from the 35-74 model show a slight decline, while from the 40-74 model they show a sharp decline starting after 1993.

B) One of the problems that CDG found for England & Wales was that the linear extrapolation gave better predictions than the log-linear, when using first part of period to "predict" the known later years. The problem of course is that linear extrapolation can lead to negative cohort values which are nonsense. However looking at the plots, it looks to me as if the log-linear estimates are too high because this model simply doesn't fit to all points past the peak - the initial fall is much gentler, hence the extrapolation starts too high with rather a discontinuity.

I therefore tried fitting a log-linear extrapolation starting from cohort 10 rather than cohort 5, this point being chosen "by eye". The resulting outputs are all labelled (2), (Figures 27,28). The extrapolations by this method were lower than the original and smoother - they started off much like the original linear extrapolation (especially for age 40-74 only) but of course don't decline so rapidly and cannot go negative. It also follows that the age-specific predictions at young ages start off similar to the original linear estimated ones.

The 40-74 std predictions are plotted as 7 black (full age range) and 8 orange (age 40-74 only). As discussed before, the extrapolation has little effect when the 25-74 model is used, so there is virtually no difference between lines 4 and 7. But for the 40-74 only model, these lower extrapolated values give lower predictions than the original line 6, and much more similar to its original linear predictions, 5, as well as the 25-74 model predictions 4 and 7.

For some other countries the "past the peak" method is hard to apply - some don't have a peak, or only a few points past (e.g. see W. Germany, Denmark above). Certainly none have anywhere near as many as England & Wales, so using less points for England & Wales may be reasonable on those grounds anyway.

C) For England & Wales and USA, the extrapolated period values are declining, but for the other countries they rise - most steeply for Greece. As such a rise may not continue (if for instance it is associated with diagnostic changes). I thought it would be interesting to look at predictions with future period effects fixed at the latest value and extrapolate as before for cohort values only. These are labelled (3). (Age-specific plots only done for Greece 30-74, Figure 29). The 40-74 std plots are shown as purple 9 (full age range) and orange a, b (age 40-74 only).

The predictions are of course lower: compare green with purple and blue with orange. This idea is rather doubtful for Greece, since when both period and cohort effects are rising, O&G is not capable of distinguishing between them.

Reviewing the work so far, the following decisions were made:

- Data below age 40 should be used where possible. There seems no good reason to omit it if it is available, as it does appear to be useful in indicating most recent trends.
- Linear extrapolation should be abandoned.

- Although the A-C model is appealing, not suffering from the redundancy problem of the full model, it does give a considerably less good fit than the full model for most countries. Although the residual correction appears to improve this when looking at the 40-74 standardised plots, the wild variations at some young age groups do not seem acceptable. The full A-P-C model will therefore be used from now on. Weighting based on 'past the peak' is too arbitrary.

4. Proposed modifications to the method

Accordingly the next step used the following method: The extrapolation of the cohort values should be based on log-linear, using weights decreasing exponentially into the past. The weights would be powers of 2 rather than of 10 since this was felt to be too heavy, virtually basing the extrapolation on the last 2 points only.

This method does however have the problem of giving the greatest weight to the last cohort values, which are the least reliably estimated. The process was therefore to be repeated omitting first one and then two final cohort values, and replacing them by values based on extrapolation of the previous values. At the same time, extrapolation of the period values would also be based on weighting by powers of 2. The three sets of results from different cohort extrapolations are labelled (4) (5) and (6), and are shown as purple c, d and e in Figures 30-35 (These are a new set of method comparison plots omitting the linear extrapolations. See also Figures 36-42, in which the dotted lines for linear extrapolation should be ignored , and Table 1).

These results suggested that omitting final cohort values was a useful exercise. For Denmark, for instance, it reflects the uncertainty about which direction the cohort values are going. It was decided that two alternatives would generally be sufficient - using all values and with one value omitted.

However it was apparent that the nature of the period extrapolation was much more important than had been previously realised. This was partly due to the fact that previous investigations had been based on England & Wales where the period pattern is very flat; also where the cohort effect is much stronger, the habit of plotting the two effects together has tended to overshadow any pattern in the period effect. Looking

afresh at the plots of period values for all 24 countries led to the hypothesis that period values may be following a common trend, with phases: level, steep rise, level, slow decline, countries with the early cohort peak more likely to be at a later part of this pattern. However, since (1) this is not based on any formal hypothesis, (2) variation between countries is considerable and (3) there is no obvious mathematical function of the required shape, extrapolation could not be based directly on this. However a log-linear extraploation based on the last two points only would always come nearer to the required shape than if previous points are also included. This is equivalent to applying the same percentage change subsequently as occurred between the final two points. By adding and subtracting a further 3% change, a range of values is demonstrated which will hopefully encompass changes in direction between phases in the period pattern.

This method for period extrapolation and the two methods for cohort extrapolation are denoted as (7) and (8). See method comparisons Figures 43-48 and age-specific Figures 49-55.

Table 2 compares the extrapolated period values by the three different methods tried: weighting by powers of 10, weighting by powers of 2 and based on last 2 points only. The "powers of 2" values generally increase most steeply (except for Denmark) and for USA they increase while the other 2 methods decrease. There is little difference between the "powers of 10" and "last two points" values, but the latter method has the advantage of defining easily understood upper and lower values. These values are not intended to represent a statistical confidence limit, but to indicate the effect on rates of a modest change in direction of the period values.

5. Finalised method

This method is now felt to be satisfactory and, to recap, is defined as:

New cohort values are estimated by log-linear extrapolation using as weights the powers of 2 decreasing into the past. The last cohort value from the fitted model should be excluded from the extrapolation procedure and replaced by an estimated value. An extrapolation based on all the cohort values should be presented as an alternative for discussion. New period values are estimated by applying the percentage change found between the last two period values to the succeeding periods. Upper and lower

estimates based on that percentage change ±3% should be presented as alternatives for discussion.

6. Results in the 6 countries

Looking at Figures 43-48, it can be seen that the general shape of the lung cancer rate is established whichever of the alternatives is accepted, although there is considerable variation in the relative size of any increase or decrease, and some variation in the position of any peak. Thus, in the USA the rates are expected to peak - the highest rate has already been reached in 1983 according to the lower period values, whereas a further 10 years increase (6%) is predicted by the upper period values. The pattern for Canada is similar although 10 years later, so it is not clear within the time span used whether the peak has been reached with the upper period values.

For England & Wales, a continuing decline is predicted, the rate in 2003 varying between 43% and 56% of the 1983 rate. Similarly in Greece a continuing increase is predicted, giving an increase over the 1983 rate of between 19% and 50%.

In W. Germany, the influence of the double cohort peak shows that the recent slight decline will not continue. But the predictions vary between virtually no change in rate to a 27% increase by 2003.

In Denmark, where there was a greater degree of uncertainty in the direction of the cohort values, and where analysis is only based from age 35, there is a more fundamental divergence of predictions. All methods agree that the rate will rise until 1993 (between 3% and 17%), but thereafter the rate may start to fall, level or continue rising, to finish 10% below or 24% above the 1983 rate.

References

1. Gooch, C.D. Forward Prediction of Lung Cancer Mortality 1981-2000. August 1984.

2. Forey, B.A. Prediction of Lung Cancer Rates 1986-2005 in 24 countries. 6 February 1989.