Appendix 2.

Codes used to define COPD and lung cancer

The codes used to define COPD and lung cancer used in these analyses is presented in Table A1 below.

Table A1. International Classification of Disease (ICD) codes used to define COPD and lung cancer

ICD5 1940-49	ICD6 1950-57	ICD7 1958-67	ICD8 1968-78	ICD9 1979-1996
Codes for COPD				
112 asthma	501 bronchitis	501 bronchitis	491 chronic bronchitis	491 chronic bronchitis
113 emphysema	502 chronic bronchitis	502 chronic bronchitis	492 emphysema	492 emphysema
	526 bronchiectasis	526 bronchiectasis	493 asthma	493 asthma
		527 emphysema	518 bronchiectasis	494 bronchiectasis
			519 chronic airways obstruction	495 extrinsic allergic alveolitis
				496 chronic airways obstruction
Codes for lung cancer				
47b Cancer of the lung and pleura	162 Malignant neoplasm of trachea, and of bronchus and lung specified as primary	162 Malignant neoplasm of bronchus and trachea, and of lung (primary)	162 Malignant neoplasm of trachea, bronchus and lung	162 Malignant neoplasm of trachea, bronchus and lung
	163 Malignant neoplasm of lung and of bronchus, unspecified as to whether primary or secondary	163 Malignant neoplasm of lung, unspecified as to whether primary or secondary		

Further information about the Bayesian age-period-cohort statistical model and its implementation in BAMP (Bayesian Age-period-cohort Modelling and Prediction) software

The statistical model used for this analysis[1] (referred to as BAMP thereafter) [1] is based on the Bayesian version of age-period-cohort model first proposed by Bezuini [2] and related to the classical age-period-cohort model described by Clayton and Schifflers [3]. The basic model used in BAMP can be expressed as

$$\eta_{ij} = \mu + \theta_i + \varphi_j + \psi_k + z_{ij} \qquad \textbf{Equation 1}$$

where

η_{ij} = log odds of death in age-group *i* and year *j*

μ = overall level

θ = age effects in age-group *i*

φ = period effects in year *j*

ψ = cohort effects in cohort *k*

z_{ij} = unstructured heterogeneity in age-group *i* and year *j* not captured by θ, φ or ψ considered to be due to unobserved observation-specific covariates

The parameter for unstructured heterogeneity (assumed to relate to unobserved covariates) can be thought of as analagous to adjustments made for overdispersion in a frequentist setting. However, within a Bayesian framework it ensures a good fit to the data.

Prior beliefs incorporated within the model

The underlying 'base' rate (parameter μ in Equation 1) is assumed to be constant and a flat uniform prior is assigned to it. BAMP uses one of two different smoothing priors defined as Random Walk 1 (RW1) and Random Walk 2 (RW2) [1]. Briefly, the RW1 prior uses first order differences of age, period or cohort parameters, favouring solution parameters with constancy i.e. it penalises more extreme solutions and assumes a smoothness of age, period and cohort trends. For example, for the age effects θ, the smoothing prior is formulated as

$$p(\theta|\kappa) \propto \exp\left[-\frac{\kappa}{2}\sum_{i=2}^{I}(\theta_i - \theta_{i-1})^2\right] \qquad \text{Equation 2}$$

where κ is a precision parameter for the age effects.

The RW2 prior uses second order differences of age, period or cohort parameters and penalises deviations from a linear trend, assuming a smoothness of the rate of change of parameters. For age effects θ, this is formulated as

$$p(\theta|\kappa) \propto \exp\left[-\frac{\kappa}{2}\sum_{i=3}^{I}(\theta_i - 2\theta_{i-1} + \theta_{i-2})^2\right] \qquad \text{Equation 3}$$

with κ again representing a precision parameter for the age effects. BAMP assumes that RW1 and RW2 priors and the unstructured heterogeneity parameter z are independent Gaussian random variables.

As with frequentist age-period-cohort models [3], a combination of very different age, period and cohort parameter sets could be responsible for the observed rates. To identify solution sets, BAMP constrains values of age, period and cohort effects to sum to zero, that is $\Sigma\theta_i = 0$, $\Sigma\varphi_j = 0$ and $\Sigma\psi_k = 0$, but does not add further arbitrary constraints to allow interpretation of linear trends as this is not necessary for making projections. As a result only non-linear trends of age, period and cohort parameters are interpretable (e.g. a change from an increase to a fall).

Frequentist age-period-cohort models[3] are typically conducted with age and period on the same time grid (typically five or ten year bands), although modifications to allow differing grids have been published [4]. Routine mortality data are often readily available by single year and five or ten year age-band and restriction to identical grids for analysis may result in loss of information. Unlike most frequentist and some other Bayesian models [5] BAMP is able to incorporate age-groups and periods on different time grids. This results in a greater overlap of birth cohorts, but this is not a problem for the Bayesian model as the model also incorporates an *a priori* assumption that consecutive cohort parameters are similar.

Implementation of the model

The model implies the posterior distribution for μ and the other unknown parameters θ, φ, ψ, and z together with their corresponding precision parameters κ, λ, υ, and δ. The model was reparameterised from z_{ij} to η_{ij} for all values of *i* and *j* following Besag et al [6] such that posterior distributions were assumed to be multivariate Gaussian. Specified values for the hyperprior settings of unknown precision parameters κ, λ, υ, and δ were chosen to be uninformative [1] (hyperpriors can be thought of as the 'starting points' or rather 'starting distributions' for iterative sampling).

Iterative sampling when using overlapping birth cohorts is too computationally intensive to be conducted using Bayesian software such as BUGS, as used for some Bayesian projections [5]. BAMP uses a block sampling algorithm[1] to allow completion of analyses within a short time-frame (typically a few minutes).

The actual running of BAMP software is straightforward. The software can be downloaded as freeware from the internet (http://www.statistik.lmu.de/sfb386/software/bamp/bamp.html last accessed 18 June 2005). Inputs required are text files for counts of numbers of deaths and population per year (we suggest at least 25 years of data, but this is not limited by the model), plus a simple text initialisation file to define parameters – a sample initialisation file available is online with the software. Results are outputted as text files that can be read by any statistical package. S-plus and R code to display results graphically is also available to download.

REFERENCES

1. Knorr-Held L, Rainer E. Prognosis of lung cancer mortality in West Germany: a case study in Bayesian prediction. Biostatistics 2001; 2(1):109-129.

2. Berzuini C, Clayton D, Bernardinelli L. Bayesian inference on the lexis diagram. Bulletin of the International Statistical Institute 1993; 50:149-164.

3. Clayton D, Schifflers E. Models for temporal variation in cancer rates. II: Age-period-cohort models. Stat Med 1987; 6:469-481.

4. Schifflers E, Smans M, Muir CS. Birth cohort analysis using irregular cross-sectional data: a technical note. Stat Med 1985; 4:63-75.

5. Bray I, Brennan P, Boffetta P. Projections of alcohol- and tobacco-related cancer mortality in central Europe. Int J Cancer 2000; 87:122-128.

6. Besag JE, Green PJ, Higdon DM, Mengersen KL. Bayesian computation and stochastic systems (with discussion). Stat Sci 1995; 10:3-66.