STATISTICS IN MEDICINE, VOL. 1, 245–259 (1982)

# AGE, PERIOD AND COHORT MODELS APPLIED TO CANCER MORTALITY RATES

C. OSMOND AND M. J. GARDNER

*MRC Environmental Epidemiology Unit, South Lab & Path Block, Southampton General Hospital, Southampton, S09 4XY, U.K.*

## SUMMARY

We develop the application of age, period and cohort models to the representation of tables of age- and period-specific rates. A derivation is given by way of a familiar graphical technique. The identifiability problem is discussed, identification techniques are reviewed and a new approach is recommended that is based upon the success of the three two-variable submodels. Other constraints are introduced that enhance interpretation. Examples are given for two sites of cancer. This approach is contrasted with other methods designed to demonstrate trends. Finally, standard errors of the parameters and tests of goodness of fit are discussed.

KEY WORDS Mortality rates Time-trends Age–period–cohort models Identification problem

## 1. INTRODUCTION

Mortality rates for a particular cause are often provided in the form of a matrix in which the rows correspond to age-groups and the columns to periods of time. This matrix is obtained from two similarly sized ones containing the numbers of deaths and person-years at risk, respectively. Our approach to the description of mortality trends is illustrated by using cancer mortality rate tables for England and Wales.[1–4] Over sixty such matrices have been formed, each for a combination of sex and site. Each has five-year age-groups and five-year periods of time. We only consider deaths between 1951 and 1980 among those under seventy. These measures are introduced to increase diagnostic reliability. The lower limit of age is decreased until a further decrease would result in an element of the matrix of rates being based upon less than twenty deaths, or an age less than fifteen years is reached. Tables I and II contain mortality rates for bladder cancer among males and lung cancer among females as examples. A larger selection of results[5] and the complete set[6] are available elsewhere.

We use the following notation:

$I$ = number of consecutive $T$-year age groups
$J$ = number of consecutive $T$-year periods
$K$ = number of cohorts $(= I + J - 1)$
$R = (r_{ij})$ is the matrix of rates
$D = (d_{ij})$ is the matrix of numbers of deaths
$Y = (y_{ij})$ is the matrix of person-years at risk

The three matrices $R$, $D$ and $Y$ are of size $I \times J$, with $r_{ij} = d_{ij}/y_{ij}$ for all $i$ and $j$.

In Section 2 traditional approaches to these data are discussed. These lead to the derivation of

0277–6715/82/030245–15$01.50
© 1982 by John Wiley & Sons, Ltd.

*Received 22 February 1982*
*Revised 26 May 1982*

age, period and cohort models contained in Section 3. The ensuing identification problem is described in Section 4, which is followed in Section 5 by an account of identification techniques used in previous work with these models. In Section 6 we specify our choice of solution. The final two sections contain results and a discussion of goodness of fit and stability of the models.

## 2. TRADITIONAL APPROACHES

Methods that have been used to describe rate matrices may be classified according to two criteria. The first is whether they provide a vector summary of the rates or represent all cells. The second is whether they describe changes indexed by period or by birth cohort. We illustrate the possibilities by discussing two approaches that will later become useful in the evaluation and derivation of our technique.

Using the data in Tables I and II we may indirectly standardize the period death rates for age with respect to the age-specific rates of 1966–1970, for example. Thus we calculate the ratio of the number of deaths observed in any five-year period to the number that would have been expected had the 1966–1970 rates applied to that period. This results in standardized mortality ratios (SMRs) for each period which are shown in the final rows of Tables I and II. For both sites the

Table I. Mortality from bladder cancer in men in England and Wales during 1951–80, ages 40–69, ICD code 188 (8th revision). Death rates per million person-years at risk. The leading diagonal corresponds to the 1910/1 cohort. Standardized mortality ratios (SMR) with respect to age-specific rates for 1966–1970.

| | | Period of death | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1951–55 | 1956–60 | 1961–65 | 1966–70 | 1971–75 | 1976–80 |
| Age at death | 40–44 | 13·6 | 16·3 | 14·4 | 12·2 | 11·4 | 9·2 |
| | 45–49 | 38·8 | 35·2 | 33·4 | 34·0 | 31·5 | 23·9 |
| | 50–54 | 85·7 | 68·9 | 70·7 | 71·3 | 66·9 | 65·0 |
| | 55–59 | 149·4 | 152·8 | 144·8 | 147·5 | 138·5 | 131·3 |
| | 60–64 | 269·3 | 264·8 | 284·5 | 276·6 | 266·2 | 244·5 |
| | 65–69 | 403·6 | 430·5 | 448·6 | 508·2 | 465·0 | 458·8 |
| | SMR | 0·94 | 0·94 | 0·96 | 1·00 | 0·94 | 0·89 |

Table II. Mortality from lung cancer in women in England and Wales during 1951–80, ages 25–69, ICD codes 162 and 163 (8th revision). Death rates per million person-years at risk. The leading diagonal corresponds to the 1925/6 cohort. Standardized mortality ratios (SMR) with respect to age-specific rates for 1966–1970.

| | | Period of death | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1951–55 | 1956–60 | 1961–65 | 1966–70 | 1971–75 | 1976–80 |
| Age at death | 25–29 | 5·5 | 4·0 | 4·3 | 3·9 | 3·9 | 2·8 |
| | 30–34 | 15·1 | 14·8 | 11·2 | 10·9 | 9·0 | 7·8 |
| | 35–39 | 28·5 | 32·1 | 32·2 | 31·4 | 26·3 | 26·5 |
| | 40–44 | 52·1 | 61·5 | 68·1 | 82·0 | 67·8 | 62·5 |
| | 45–49 | 88·7 | 105·5 | 137·3 | 156·6 | 183·0 | 159·5 |
| | 50–54 | 139·2 | 170·5 | 218·2 | 285·8 | 330·7 | 360·3 |
| | 55–59 | 206·3 | 243·2 | 309·7 | 402·5 | 499·0 | 570·3 |
| | 60–64 | 287·7 | 332·7 | 424·0 | 519·4 | 676·7 | 857·9 |
| | 65–69 | 356·6 | 384·3 | 518·5 | 661·9 | 816·1 | 1030·5 |
| | SMR | 0·55 | 0·63 | 0·80 | 1·00 | 1·21 | 1·43 |

SMRs have a similar pattern to the rates in the oldest age-groups. These age-groups have the highest rates, and tend to dominate the SMR. The pattern should not, therefore, be interpreted as an indication that rates at all ages behave similarly. We need to consider the behaviour in separate age-groups. Other choices of age-range for standardization will not yield a consistent pattern.

Another traditional approach to these data has been to plot the rates against age, joining all points on a corresponding cohort.[7,8] Such points appear along descending diagonals of the rate matrix. For many sites it has been observed that cancer mortality rates increase roughly as a power of age at death,[9] so that a doubly logarithmic scale is natural. In Figures 1 and 2 the results of this process are shown for the data in Tables I and II, alternate cohorts being plotted for clarity. In both cases the rates may be seen to increase to a maximum for particular cohorts and subsequently decline. Only a few rates contribute to the earliest and latest cohorts, and for recent cohorts these are for young people and are therefore based upon smaller numbers of deaths.

Standardization collapses the rate matrix down its columns to provide a period-based summary vector. This may also be done down the diagonals (cohorts) to produce a standardized cohort mortality ratio (SCMR).[10] SCMRs are shown on Figures 1 and 2. Summary vectors may be constructed in even more complicated ways.[11] The display of rates (originally due to Case) represents all available data points and emphasizes the importance of cohort. Other combinations are possible. In all, six Case-like displays are available if we plot rate (or log rate) against one of age, period or cohort and then join points corresponding to one of the others. Also, we may display all the rate values by considering the rates as forming a surface and making a perspective plot, thoughtfully choosing the eye position.[12]

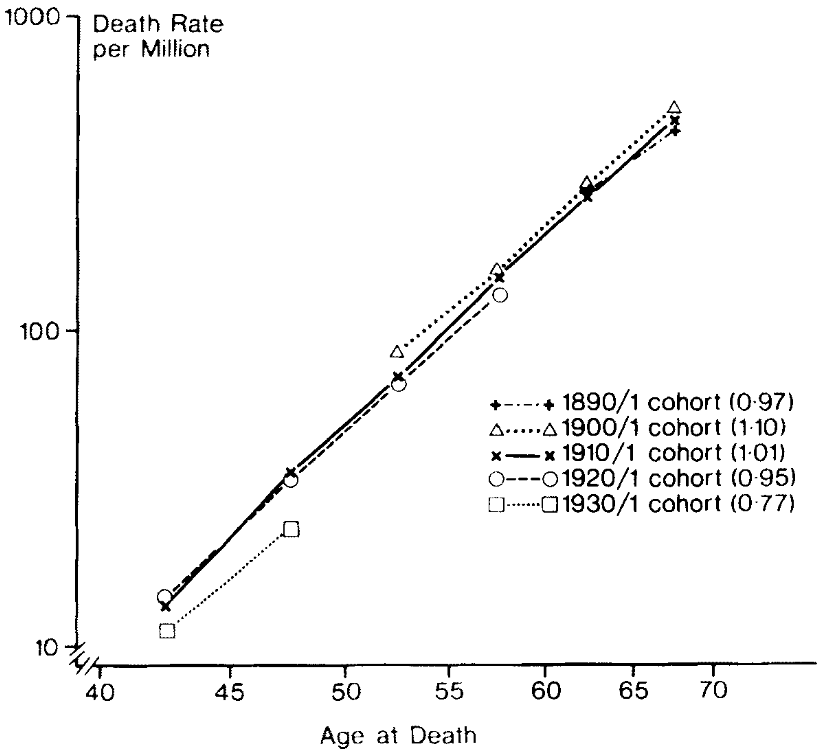

Figure 1. Mortality from bladder cancer in men in England and Wales during 1951–80, ages 40–69, ICD code 188 (8th revision). Death rates per million are plotted against age at death on a doubly logarithmic scale. Points corresponding to the same cohort are joined, but only alternate cohorts are plotted for clarity. Standardized cohort mortality ratios are included in brackets.

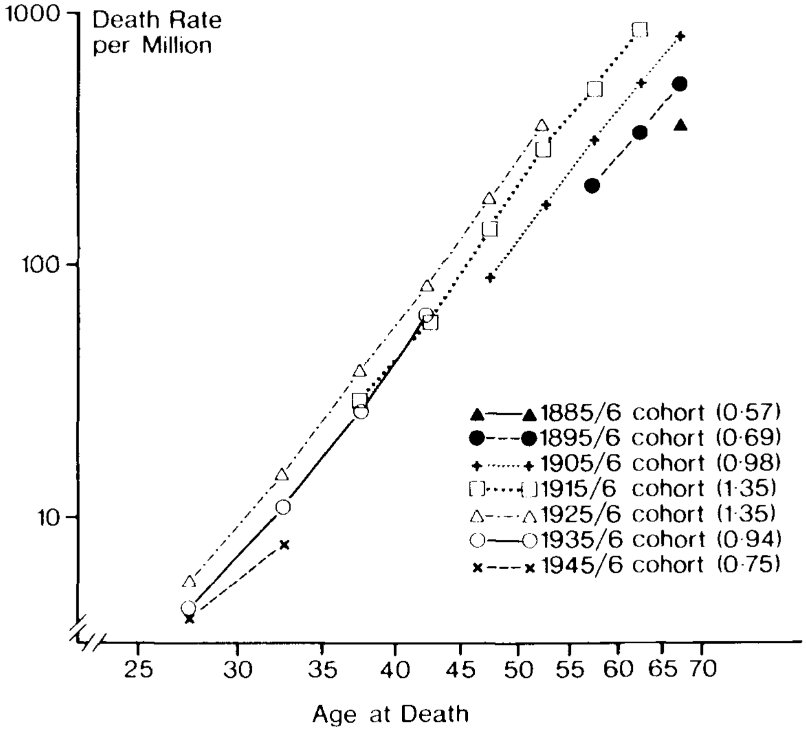

Figure 2. Mortality from lung cancer in women in England and Wales during 1951–80, ages 25–69, ICD codes 162 and 163 (8th revision). Graph produced as for Figure 1.

In what follows we have two main aims in developing the method. Firstly, we seek to describe the simultaneous effect of period- and cohort-indexed factors rather than just one set. Secondly, we seek to simplify, but not oversimplify, the available mass of data. The final result is designed to have a ready interpretation and visual simplicity.

## 3. AGE, PERIOD AND COHORT MODELS

Figures 1 and 2 are the starting point for the derivation of these models. Within each Figure the cohort curves may be seen to have a similar shape. The effect of differing cohort is to shift the entire curve up or down. In these cases the cohort curves are extremely straight, but for other sites this is not so. Thus generally the entire set of rates may be summarized by a curve and a set of shifts. The curve is determined by a set of values corresponding to age; the shifts correspond to cohorts. Effectively we are fitting the model

$$\log r_{ij} = \log a_i + \log c_{I-i+j} \tag{1}$$

where the $\log a_i$ terms describe the curve and the $\log c_{I-i+j}$ terms are the shifts. $I-i+j$ is the cohort corresponding to age-group $i$ and period $j$, the cohorts being numbered from the earliest in time. The only remaining requirement is the specification of an origin for measuring the shifts. Once this is provided, a unique solution is available. Exponentiating (1), the rate is seen to be approximated as a product of $a_i$ and $c_{I-i+j}$ terms, which we refer to as the age values and cohort values, respectively.

It is quite possible that there will also be an influence on the rates due to factors indexed by period of death. Improvements in diagnostic technique might occur at a point in time and affect recorded mortality rates at all ages. Treatment improvements could act similarly, as could other environmental influences. A set of period of death values, $p_j$, may be fitted in a similar way. Although the three variables age, period and cohort are not independent, they may each exert a simultaneous influence in that they index contributory causal factors.

We adopt a general least squares formulation for this problem of which (1) may be seen to be a particular case. First we define the residual function, $f$, by

$$f(\mathbf{a}, \mathbf{p}, \mathbf{c}) = \sum_{i,j} d_{ij}(\log r_{ij} - \log a_i - \log p_j - \log c_{I-i+j})^2 \tag{2}$$

The weights $d_{ij}$ are the deaths, as before, and are chosen because they are inversely proportional to the sampling variances of the logarithms of the rates,[13] where the number of deaths is considered as a Poisson random variable. Partial differentiation with respect to $\log a_i$, $\log p_j$ and $\log c_k$ produces a set of $I + J + K$ normal equations in the $I + J + K$ variables. However three of the equations are redundant as they are re-expressions of the others. Two additional equations may be provided by the need for two origin fixes. One further equation is needed and this is the source of the identification problem discussed in Section 4. Before that we consider the origin fixes.

At the minimum of $f$ we find that

$$\sum_j \log p_j\left(\sum_i d_{ij}\right) + \sum_k \log c_k\left(\sum_{\{(i,j):I-i+j=k\}} d_{ij}\right) = \sum_{i,j}(\log r_{ij} - \log a_i)d_{ij} \tag{3}$$

This relationship suggests three possible constraints, of which any two may be taken to provide an origin fix, forcing the other to be satisfied at the minimum.

$$\text{ConA:} \quad \sum_{i,j}(\log r_{ij} - \log a_i)d_{ij} = 0 \tag{4}$$

$$\text{ConP:} \quad \sum_j \log p_j\left(\sum_i d_{ij}\right) = 0 \tag{5}$$

$$\text{ConC:} \quad \sum_k \log c_k\left(\sum_{\{(i,j):I-i+j=k\}} d_{ij}\right) = 0 \tag{6}$$

If ConA is satisfied the age values may resemble the corresponding age-specific rates. If ConP is satisfied a typical period value is unity. If ConC is satisfied a typical cohort value is unity. Seven models may then be defined.

*Age model*: Given $\mathbf{p}_0$ satisfying ConP and $\mathbf{c}_0$ satisfying ConC minimize $f(\mathbf{a}, \mathbf{p}_0, \mathbf{c}_0)$ over $\mathbf{a}$. ConA will be satisfied at the minimum.

*Period model*: Given $\mathbf{c}_0$ satisfying ConC and $\mathbf{a}_0$ satisfying ConA minimize $f(\mathbf{a}_0, \mathbf{p}, \mathbf{c}_0)$ over $\mathbf{p}$. ConP will be satisfied at the minimum.

*Cohort model*: Given $\mathbf{a}_0$ satisfying ConA and $\mathbf{p}_0$ satisfying ConP minimize $f(\mathbf{a}_0, \mathbf{p}_0, \mathbf{c})$ over $\mathbf{c}$. ConC will be satisfied at the minimum.

*Period and cohort model*: Given $\mathbf{a}_0$ satisfying ConA minimize $f(\mathbf{a}_0, \mathbf{p}, \mathbf{c})$ over $\mathbf{p}$ and $\mathbf{c}$. Requiring ConP at the minimum ensures ConC. Requiring ConC ensures ConP. The solutions are the same.

*Age and cohort model*: Given $\mathbf{p}_0$ satisfying ConP minimize $f(\mathbf{a}, \mathbf{p}_0, \mathbf{c})$ over $\mathbf{c}$ and $\mathbf{a}$. Requiring ConC at the minimum ensures ConA. Requiring ConA ensures ConC. The solutions are the same.

*Age and period model*: Given $\mathbf{c}_0$ satisfying ConC minimize $f(\mathbf{a}, \mathbf{p}, \mathbf{c}_0)$ over $\mathbf{a}$ and $\mathbf{p}$. Requiring ConA at the minimum ensures ConP. Requiring ConP ensures ConA. The solutions are the same.

*Age, period and cohort model*: Minimize $f(\mathbf{a}, \mathbf{p}, \mathbf{c})$ over $\mathbf{a}$, $\mathbf{p}$ and $\mathbf{c}$. Requiring any two of ConA, ConP and ConC at the minimum ensures the other with the same solution. However extra requirements are needed to resolve the identification problem.

All of these models are additive on the log scale, multiplicative on the arithmetic scale. The simplest $\mathbf{a}_0$ that satisfies ConA is given by

$$\log a_{0i} = \left(\sum_j d_{ij} \log r_{ij}\right) \Big/ \left(\sum_j d_{ij}\right) \tag{7}$$

We suggest two possible choices for $\mathbf{p}_0$ and $\mathbf{c}_0$. In the null case in which we want to disregard completely variation indexed by period we may use $\mathbf{p}_0 = \mathbf{1}_J$ (the $J$-vector of ones). Correspondingly we may use $\mathbf{c}_0 = \mathbf{1}_K$. As an alternative we may make an allowance for some period indexed variation by using, for example, the SMR. Although this will not satisfy ConP we may find a vector proportional to it which does so. Similarly we may make components of $\mathbf{c}_0$ proportional to the SCMR to satisfy ConC.

Each solution from one of the submodels may be regarded as a point in Euclidean space of dimension $I+J+K$ denoted by $\mathbf{R}^{I+J+K}$. All considerations of solutions as points in Euclidean space will refer to the logarithms of the sets of values and not the values themselves. The lack of identifiability of the full three variable model causes solutions satisfying the constraints to lie on a straight line in $\mathbf{R}^{I+J+K}$.

## 4. THE IDENTIFICATION PROBLEM

In this section we lay aside the requirement that solutions of the full three variable model should satisfy the constraints. This is done purely for notational and illustrative convenience.

Let $(\mathbf{a}, \mathbf{p}, \mathbf{c})$ be a minimum of $f$ for the full model. Another minimum is given by $(\mathbf{a}', \mathbf{p}', \mathbf{c}')$ where

$$\begin{aligned} \log a_i' &= \log a_i + \mu + \lambda(I-i) \\ \log p_j' &= \log p_j + \nu + \lambda j \\ \log c_k' &= \log c_k - \mu - \nu - \lambda k \end{aligned} \tag{8}$$

The summation of $\log a_i'$, $\log p_j'$ and $\log c_k'$ corresponding to a particular rate $(r_{ij})$ causes the cancellation of all terms in $\lambda$, $\mu$ and $\nu$. For $\mu$ and $\nu$ the cancellation is immediate. These two parameters correspond to the fixing of origins. For $\lambda$ the cancellation corresponds to the relation

$$k = I - i + j \tag{9}$$

which arises from the intrinsic dependence of age, period and cohort. When the residual function, $f$, is unweighted and corresponding constraints are applied, the midpoint of each set of logarithms of values becomes fixed. In this case the $\lambda$ terms produce extra shifts proportional to the distance from the midpoint. For small changes in $\lambda$ these look like rotations. The signs in (8) force the age and cohort values to rotate in an opposite direction to the period values. The introduction of weighting frees the midpoint but the effect of a small change in $\lambda$ is similar in practice.

To illustrate the problem, consider the following rate matrix.

$$R = \begin{pmatrix} \alpha & \alpha\theta & \alpha\theta^2 \\ \beta & \beta\theta & \beta\theta^2 \\ \gamma & \gamma\theta & \gamma\theta^2 \end{pmatrix} \tag{10}$$

One possible identification (which fits the rates perfectly) is

$$\begin{array}{ll} \text{age values:} & \alpha,\ \beta,\ \gamma \\ \text{period values:} & 1,\ \theta,\ \theta^2 \\ \text{cohort values:} & 1,\ 1,\ 1,\ 1,\ 1 \end{array} \tag{11}$$

Another is

$$\begin{array}{ll} \text{age values:} & \alpha,\ \beta\theta,\ \gamma\theta^2 \\ \text{period values:} & 1,\ 1,\ 1 \\ \text{cohort values:} & \theta^{-2},\ \theta^{-1},\ 1,\ \theta,\ \theta^2 \end{array} \qquad (12)$$

The first identification seems to imply no variation indexed by cohort. The second seems to imply no variation indexed by period. Indeed, intermediate or even more extreme solutions may be generated with equal facility. Thus no internal means may be devised for the estimation of the gradients in the sets of values. First differences are inestimable (via $\lambda$) and gradients are arbitrary. The logical dependence between age, period and cohort is the root of this problem. All that would seem justified would be the statement that the rates are increasing ($\theta > 1$), stable ($\theta = 1$) or decreasing ($\theta < 1$).

Certain features are independent of identification. For example any differences higher than first order are estimable because $\lambda$ terms cancel. Also some estimates of future rates are not changed. Thus in our example the period and cohort-values are linear on the log scale for all identifications. Extending them linearly and recombining produces a set of rate estimates ($\alpha\theta^3$, $\beta\theta^3$, $\gamma\theta^3$ for the next period etc.) that are independent of identification.

Before discussing our approach to this problem we provide a brief review of other treatments of this form of modelling, considering their solution to the identification problem.

## 5. PREVIOUS USE OF AGE, PERIOD AND COHORT MODELS

The earliest use of this form of modelling was due to Kermack *et al.* in 1934.[14] They used an age and cohort model to study all-cause death rates. This approach was extended by others to treat tuberculosis,[15] breast cancer[16] and a variety of other sites.[17] These uses were descriptively successful, had no identification problem and demonstrated that a submodel may often be adequate.

Barrett has published several papers[13, 18–21] treating cancer mortality for a particular site in an epidemiological context by using the full model. He obtains particular identifications by arbitrarily assigning value zero to two parameters, and equating a pair. Clearly the choice of parameters can drastically affect the appearance of the sets of values. Grouping sets of parameters is equivalent to equating them and produces similar results. Barrett recognizes the problems and recommends looking for patterns that go beyond linear trends. In common with Price[22] features such as peaks are sought. Yet even these are identification dependent.

Another approach has been to specify the distributional form of one of the sets of values. The form chosen is often derived from the appearance of the rates. Success is dependent upon the appropriateness of the distribution. Some strong prior suggestion is needed. Greenberg *et al.*[23] analyse syphilis incidence by constraining the age values to have a Pearson Type III distribution to allow for an early peak and rapid decline. Beard[24] uses the distribution of age-specific rates at a point in time for his age values and a measure of cohort cigarette consumption for his cohort values, in analysing lung cancer. Day[25] considers rate matrices for the same site but for different cancer registries and constrains the age values in the different registries to be the same. Finally Feinberg and Mason,[26] treating measures of educational attainment, equate age values for six consecutive groups, arguing that once a certain age has been reached advances in educational attainment become negligible. All four of these papers involve some external criteria to provide an identification. Meaningful results are dependent upon the suitability of the criteria. Day is able to perform conventional tests of the assumption of equal age values, so internal checking is available in his approach. It is not likely that the assumption would hold for different sites, and this analysis cannot be used for just one table of rates.

Others have seen the existence of a linear trend in a set of values (or their logarithms) as an indication that the correct identification should remove this as far as possible. The justification for this has been that stable patterns of values would be represented as such, whereas previously they had been seen as systematically changing. Sacher[27] achieved this for cohort values by forcing the regression line through the first six of them to be flat. Holman *et al.*[28] equate the first and last period values to eliminate any linear trend, but then argue against the existence of significant period values. Pullum[29] adopts a more sophisticated approach to the removal of linear trend. He measures how far each of the three sets of logarithms of values departs from linearity. The nearer a set to linearity, the less trend that set is given in the final identification. The simple example of Section 4 highlights the problems associated with this technique. Small perturbations of the rate matrix could result in large changes in identification. Robustness of the identification must be an important consideration.

James and Segal[30,31] develop two models in which an age/period interaction is introduced. Specifically

$$\log r_{ij} = \alpha_i \delta_j + \beta_j + \gamma_k \tag{13}$$

and the alternative in which an age term instead of a period term stands by itself. These models do not suffer from lack of identifiability, but this must be balanced against an increase in descriptive complexity and computational effort required to attain the minimum, the need to justify the form of model chosen and the instability of the parameter estimates under certain conditions. For example, in (13), if either all $\delta_j$ values are equal or the $\alpha_i$ values are linear then no unique solution is available and the parameters are highly unstable. For many cancer sites the logarithms of rates are approximately linear with logarithm of age so that this may often be a genuine difficulty.

Much of the recent use of age, period and cohort modelling has been in demography. A recent debate[32-37] in this field has centred upon both the practical[33] and logical[36] foundation of the method. Hobcraft and Gilks[37] have attempted to resolve the latter problem by demonstrating that although the variables are dependent, there may well exist influences that may be indexed by each of them, and that act independently to produce the observed mortality rates. Their approach to the question of the appropriateness of the additivity assumption is referred to in Section 8.

## 6. CHOICE OF IDENTIFICATION

Solutions for any of the submodels referred to in Section 3 may be obtained by a direct matrix inversion of the set of normal equations with constraints. For the age and period models, given $\mathbf{c}_0$ satisfying ConC let the unique minimum of $f(\mathbf{a}, \mathbf{p}, \mathbf{c}_0)$ be located at $\mathbf{X}_c = (\hat{\mathbf{a}}, \hat{\mathbf{p}}, \mathbf{c}_0)$. Correspondingly we define $\mathbf{X}_p$ for the age and cohort model, $\mathbf{X}_a$ for the period and cohort model. For each of these solutions we may measure the goodness of fit by the mean residual sum of squares. Thus

$$\begin{aligned} \rho_c &= \frac{1}{(I-1)(J-1)} f(\mathbf{X}_c) \\ \rho_p &= \frac{1}{(I-1)(J-2)} f(\mathbf{X}_p) \\ \rho_a &= \frac{1}{(I-2)(J-1)} f(\mathbf{X}_a) \end{aligned} \tag{14}$$

define the mean residual sum of squares for the two variable submodels. The smaller these values, the better the fit. None is likely to be zero in practice, but if one is, perfect fit has been obtained. Indeed it is sometimes possible to explain the variation quite adequately using just two variables.

Solutions to the full model are parametrized by $\lambda$. We denote solutions by $\mathbf{X}(\lambda)$ and the mean residual sum of squares by $\rho$ where

$$\rho = \frac{1}{(I-2)(J-2)} f(\mathbf{X}(\lambda)) \tag{15}$$

is independent of $\lambda$.

How different are the two variable solutions from the set of possible three variable solutions that is parametrized by $\lambda$? We may use Euclidean distance in $\mathbf{R}^{I+J+K}$ to measure this, which we shall denote by $\|\cdot\|$. This is the square root of the sum of all squared co-ordinate differences and represents the natural extension of the concept of distance in lower dimensional spaces. As always the points in Euclidean space have the logarithms of the sets of values as their co-ordinates, we then define

$$\begin{aligned} d_c(\lambda) &= \|\mathbf{X}_c - \mathbf{X}(\lambda)\| \\ d_p(\lambda) &= \|\mathbf{X}_p - \mathbf{X}(\lambda)\| \\ d_a(\lambda) &= \|\mathbf{X}_a - \mathbf{X}(\lambda)\| \end{aligned} \tag{16}$$

Each of these is a convex quadratic function of $\lambda$ that is minimized by the projection of the two-variable solution onto the line of three-variable solutions. If we were to select the point reached by projection from a single two-variable model we would be in effect constraining the other variable to show no overall linear trend. Thus in the simple example of Section 4, the age and period model solution $\mathbf{X}_c$ based on $\mathbf{c}_0 = \mathbf{1}_5$ is actually contained in the line of full model solutions and allows no trend in cohort values, since projection is not needed.

The example suggests the use of an intermediate value for $\lambda$ so that at least both sets of period and cohort values should increase for $\theta > 1$. The only valid conclusion in that case seemed to be that the rates were increasing. Allowing a set of period or cohort values to decrease would be misleading. Of course it is possible that increasing mortality is the net result of (say) decreasing period-indexed factors and increasing cohort-indexed factors. Very often it is the period and cohort values that are of primary interest. Thus it is useful to incorporate the solution $\mathbf{X}_a$ into the weighting procedure we suggest for selecting a particular $\lambda$. To obtain a specific solution we weight the distances inversely by the mean residual sums of squares and minimize

$$g(\lambda) = \frac{d_c(\lambda)}{\rho_c} + \frac{d_p(\lambda)}{\rho_p} + \frac{d_a(\lambda)}{\rho_a} \tag{17}$$

This is one of the many measures that could be suggested. It has the advantage that it is a simple convex quadratic in $\lambda$ with a unique minimum. Differentiation of $g(\lambda)$ with respect to $\lambda$ produces a linear equation in the $I+J+K$ parameters, which must be satisfied at the minimum. This brings the set of normal equations with constraints to full rank for an immediate matrix inversion. The computational requirements for solution may also be met by a package such as GLIM which may be used to provide standard errors and goodness of fit tests.

The minimization of $g(\lambda)$ should never provide values with an unhelpful over-emphasis on one particular variable. If the rates are increasing (or decreasing) across the whole table, this will be reflected in both period and cohort values. If small changes are introduced into the rate matrix these will be reflected in small changes in the parameter values. However the particular values must always be treated with some circumspection. Our confidence in them will vary according to how greatly the $\lambda$-values that minimize $d_c(\lambda)$, $d_p(\lambda)$ and $d_a(\lambda)$ are spread. The whole problem is analogous to that of deciding whether to measure the age distribution of a disease cross-sectionally or along cohorts. These two processes produce different results that have different gradient (i.e. $\lambda$) properties. The extent of this difference then corresponds to the difference in $\lambda$s. Here we are taking a weighted mixture of the available approaches.

In the results that we present in Section 7 we use the three two-variable submodels based upon

$$\mathbf{c}_0 = \mathbf{1}_K \text{ (the null case)}$$

$$\mathbf{p}_0 = \mathbf{1}_J \text{ (the null case)} \qquad (18)$$

and

$$\mathbf{a}_0 \text{ as defined in (7)}$$

This last choice is the simplest available, although derived from a period-based point of view.

## 7. EXAMPLES

Figures 3 and 4 contain the results of this procedure when applied to bladder cancer in men and lung cancer in women respectively. Both figures contain two graphs. On the left graph the age value is plotted against age on a doubly logarithmic scale. On the right hand graph calendar year is used for the $x$-axis, since this is the variable by which both birth cohort and period of death are measured. Thus two lines appear on this graph, one plotting cohort values, the other period values. The typical value of unity is represented by a horizontal dashed line. In neither of these examples does period of death contribute as much as cohort, although this is not always the case. For example, cancer of the oesophagus shows recent increases corresponding to period of death.[5] Here both sites have cohort values that rise to a distinct peak; bladder cancer in men in 1900/1, lung cancer in women in 1925/6.

Parameter values obtained for each two variable model and the three variable model are

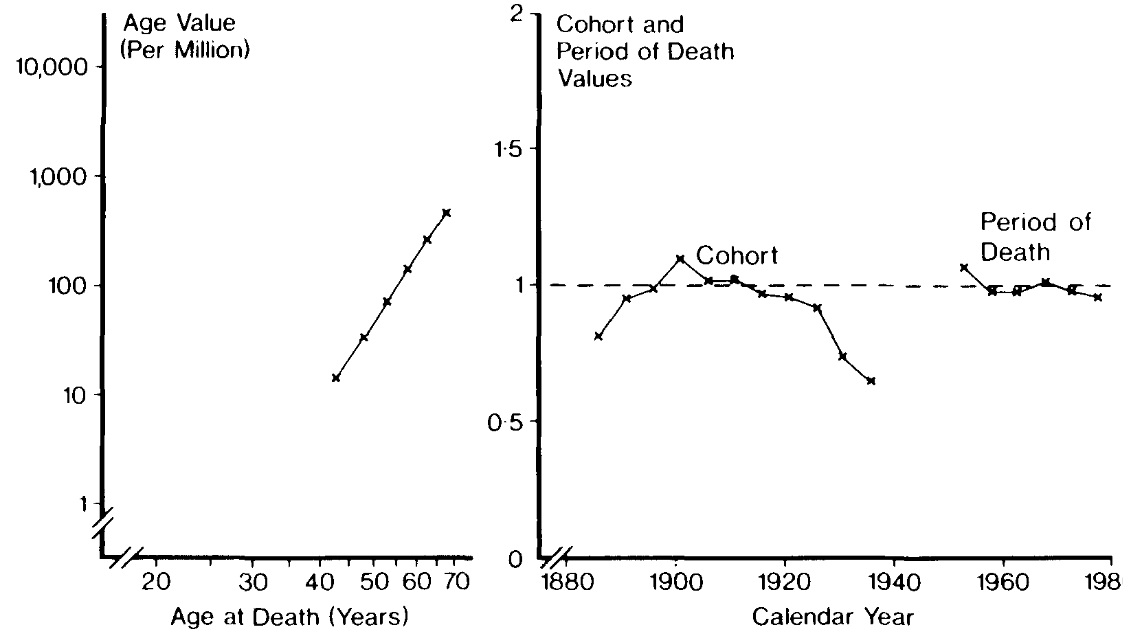

Figure 3. Mortality from bladder cancer in men in England and Wales during 1951–80, ages 40–69, ICD code 188 (8th revision). Age, cohort and period of death values (see page 256) are plotted.

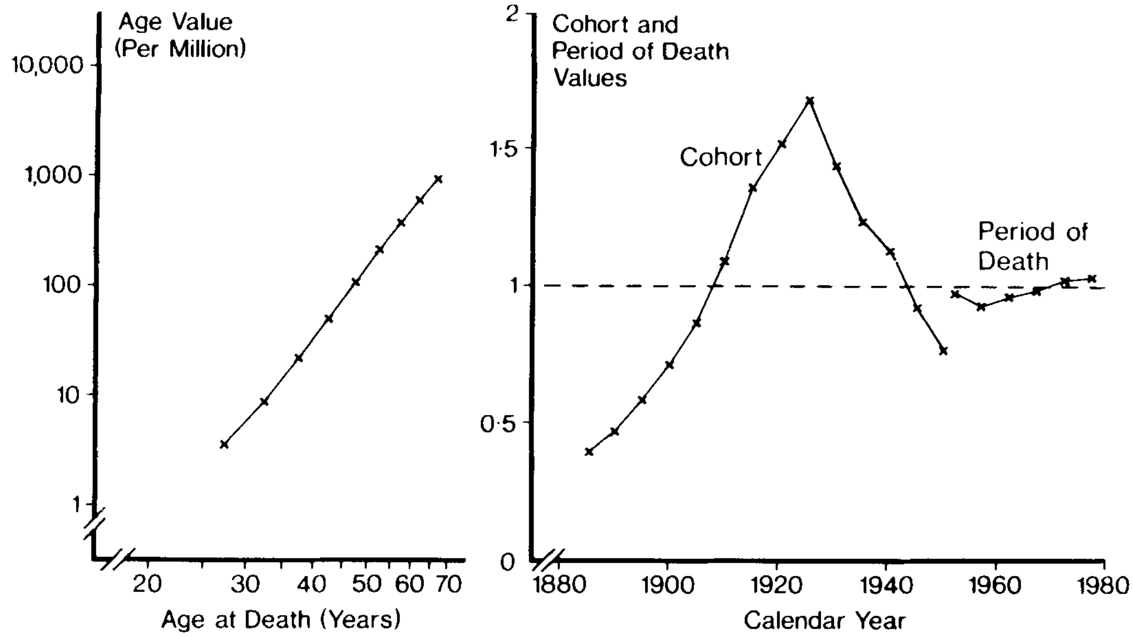

Figure 4. Mortality from lung cancer in women in England and Wales during 1951–80, ages 25–69, ICD codes 162 and 163 (8th revision). Age, cohort and period of death values (see page 257) are plotted.

provided in Tables III and IV. Also included in these tables are adjusted SMRs and SCMRs for comparison, and mean residual sums of squares. The adjustments force the SMR and SCMR to satisfy ConP and ConC, and so have an 'average' value of unity. Several points arise. First, it should be noted that the age and period model and the age and cohort model can produce age values that are markedly different. These correspond to measuring cross-sectionally and along cohorts respectively, as discussed in Section 6. Secondly there is a strong similarity between the adjusted SMR and the period values from the age and period model. In both cases the greatest difference is 0·01. Thus using adjusted SMR as $\mathbf{p}_0$ in the age and cohort model has some justification, for it almost corresponds to the first step in an iterative projection method which eventually converges to a solution of the full model. Using $\hat{\mathbf{c}}$ from that age and cohort model as $\mathbf{c}_0$ in the age and period model we may further reduce $f$, and so on. Other choices of starting vectors and models are available to provide such an iterative technique. The minimization of $g(\lambda)$, as in Section 6, provides a more direct projection. Thirdly, there is not such a clear correspondence between the adjusted SCMR and the cohort values from the age and cohort model. Since there will not always be a cohort encompassing all age-groups it is usual practice to take average age-specific rates as standard. This approach is period-based and is inappropriate when a strong cohort factor exists, as in these cases. Thus a comparison of rates in the 1915/6 and 1935/6 cohorts for common age-groups for female lung cancer suggests similarity of value, consistent with the age and cohort model rather than the SCMR. The vector $\mathbf{a}_0$ defined in (7) is likewise period-based and inclined to cause underestimates of strong cohort effects in the period and cohort model. Fourthly, the full model solution may be seen to be weighted towards the best fitting submodel. For bladder cancer in men this is the period and cohort model; for lung cancer in women it is the age and cohort model. Fifthly it is interesting to note that the age values produce a very straight line on the doubly logarithmic plots of Figures 3 and 4. Linearity is not invariant under identification. This observation may be seen to add further weight to either the choice of identification or the power law for cancer rates. If one of these is accepted the other is given more justification.

Table III. Mortality from bladder cancer in men in England and Wales during 1951–80, ages 40–69, ICD code 188 (8th revision): SMR; SCMR; age, period and cohort models.

| | | SMR | SCMR | A&P | A&C | P&C | A, P&C |
|---|---|---|---|---|---|---|---|
| | 40–44 | | | 13·1 | 15·3 | | 14·5 |
| | 45–49 | | | 33·2 | 36·1 | | 34·7 |
| Age at death | 50–54 | | | 71·6 | 72·8 | | 71·0 |
| | 55–59 | | | 143·9 | 143·7 | | 142·0 |
| | 60–64 | | | 267·7 | 264·4 | | 264·5 |
| | 65–69 | | | 457·3 | 452·1 | | 460·4 |
| | 1951–55 | 1·00 | | 1·00 | | 1·07 | 1·07 |
| Period of death | 1956–60 | 1·00 | | 0·99 | | 1·00 | 0·99 |
| | 1961–65 | 1·02 | | 1·01 | | 0·99 | 0·98 |
| | 1966–70 | 1·06 | | 1·06 | | 1·02 | 1·01 |
| | 1971–75 | 0·99 | | 0·99 | | 0·98 | 0·99 |
| | 1976–80 | 0·94 | | 0·94 | | 0·96 | 0·96 |
| | 1885/6 | | 0·89 | | 0·89 | 0·82 | 0·82 |
| | 1890/1 | | 0·97 | | 0·98 | 0·94 | 0·95 |
| | 1895/6 | | 1·00 | | 1·00 | 0·99 | 0·99 |
| | 1900/1 | | 1·10 | | 1·10 | 1·09 | 1·09 |
| | 1905/6 | | 1·02 | | 1·03 | 1·02 | 1·02 |
| Cohort | 1910/1 | | 1·01 | | 1·01 | 1·03 | 1·02 |
| | 1915/6 | | 0·95 | | 0·95 | 0·98 | 0·98 |
| | 1920/1 | | 0·94 | | 0·92 | 0·97 | 0·96 |
| | 1925/6 | | 0·92 | | 0·88 | 0.95 | 0·93 |
| | 1930/1 | | 0·77 | | 0·69 | 0·79 | 0·74 |
| | 1935/6 | | 0·71 | | 0·60 | 0·73 | 0·66 |
| Mean residual sum of squares | | | | 0·954 | 0·231 | 0·203 | 0·137 |

Finally, we may compare the mean residual sums of squares. In neither case does the age and period model fit the data as well as the age and cohort model. This is reflected in the full model solutions which, although improving the mean residual sums of squares, do not have period of death values greatly different from unity.

The overall suggestions are of hazards that have affected different generations to a different degree, and have been disappearing from the environment since their influence struck the maximal cohorts. The peak for bladder cancer in men probably corresponds to those who suffered the highest industrial exposures to aromatic amines which are known to be an important cause of this tumour.[38] However the whole curve, including the timing of the peak, resembles that for lung cancer in men.[6] Smoking is also known to be associated with bladder cancer so that this would add more force to the 1900/1 maximum, as this is also the maximal cohort for lung cancer in men. The decline in more recent cohorts could be due to both improved working conditions and safer smoking habits.

That women started smoking later than men is reflected in the later position of the peak cohort for lung cancer, 1925/6 rather than 1900/1. Numbers of cigarettes smoked by successive generations of either sex have not declined to any great extent,[39] raising the question as to what has caused lung cancer decreases. Reduction of tar content of cigarettes has been suggested,[40] but not unanimously accepted.[41] Alternatively reductions of air pollution may have been important.[42]

Table IV. Mortality from lung cancer in women in England and Wales during 1951–80, ages 25–69, ICD codes 162 and 163 (8th revision): SMR; SCMR; age, period and cohort models.

| | | SMR | SCMR | A&P | A&C | P&C | A, P&C |
|---|---|---|---|---|---|---|---|
| Age at death | 25–29 | | | 4·8 | 3·1 | | 3·5 |
| | 30–34 | | | 14·5 | 8·0 | | 8·9 |
| | 35–39 | | | 34·7 | 19·6 | | 21·4 |
| | 40–44 | | | 74·4 | 44·5 | | 47·5 |
| | 45–49 | | | 150·3 | 99·6 | | 104·3 |
| | 50–54 | | | 267·6 | 201·7 | | 207·0 |
| | 55–59 | | | 396·3 | 359·0 | | 361·0 |
| | 60–64 | | | 551·0 | 613·4 | | 604·5 |
| | 65–69 | | | 672·4 | 935·7 | | 906·8 |
| Period of death | 1951–55 | 0·55 | | 0·55 | | 0·65 | 0.98 |
| | 1956–60 | 0·63 | | 0·63 | | 0·71 | 0·93 |
| | 1961–65 | 0·80 | | 0·80 | | 0·84 | 0·97 |
| | 1966–70 | 1·00 | | 1·00 | | 0·99 | 1·00 |
| | 1971–75 | 1·22 | | 1·21 | | 1·15 | 1·02 |
| | 1976–80 | 1·43 | | 1·44 | | 1·32 | 1·04 |
| Cohort | 1885/6 | | 0·54 | | 0·38 | 0·78 | 0·40 |
| | 1890/1 | | 0·56 | | 0·44 | 0·78 | 0·47 |
| | 1895/6 | | 0·66 | | 0·55 | 0·85 | 0·59 |
| | 1900/1 | | 0·78 | | 0·70 | 0·91 | 0·72 |
| | 1905/6 | | 0·93 | | 0·86 | 0·96 | 0·87 |
| | 1910/1 | | 1·18 | | 1·10 | 1·07 | 1·10 |
| | 1915/6 | | 1·28 | | 1·40 | 1·14 | 1·37 |
| | 1920/1 | | 1·28 | | 1·60 | 1·12 | 1·53 |
| | 1925/6 | | 1·29 | | 1·80 | 1·12 | 1·69 |
| | 1930/1 | | 1·04 | | 1·57 | 0·90 | 1·44 |
| | 1935/6 | | 0·89 | | 1·38 | 0·77 | 1·25 |
| | 1940/1 | | 0·83 | | 1·29 | 0·70 | 1·14 |
| | 1945/6 | | 0·72 | | 1·07 | 0·59 | 0·93 |
| | 1950/1 | | 0·65 | | 0·90 | 0·51 | 0·77 |
| Mean residual sum of squares | | | | 4·565 | 0·284 | 1·792 | 0·172 |

This form of analysis has provided a ready, visual summary of the rate matrix in terms of the three variables, age, period and cohort. In this respect it is more informative than standardization analyses that concentrate upon one variable. Yet it does represent a summary of the set of rates. As a descriptive tool it enhances interpretation of the trends in disease mortality (or incidence) as has been seen from the two examples. Any interpretation must be made in the awareness of the problems caused by lack of identifiability. We have chosen an identification that is designed to produce solutions that are intermediate in that they apportion variation between projected two variable solutions. Small changes in the inestimable parameter $\lambda$ result in what look like rotations, with age and cohort having the opposite direction of rotation to period.

## 8. GOODNESS OF FIT

It is possible to test the goodness of fit of these models in the conventional way. Tests and standard errors of the individual parameters may be obtained from GLIM if required. Hobcraft and Gilks[37] have shown that there is a nested sequence of models that permits various tests of the additivity assumption implicit in this formulation. This is done with respect to models that allow quadratic and cubic interaction terms in age and period.

We mention one other procedure that is useful in assessing the variability of the extreme cohort values. New versions of the matrix, $D$, of numbers of deaths may be obtained by assuming that their elements are realizations of independent Poisson distributions with mean equal to the observed value. The resulting rate matrices are treated as before. Plotting envelopes of the values from several such realizations gives an indication of their variability. Generally the last cohort values are most variable because they are based upon small numbers of deaths. Cohort values near the centre are the least varying.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge helpful discussions with Professor E. D. Acheson, Dr. A. M. Adelstein and Mr. W. R. Gilks.

## REFERENCES

1. Office of Population Censuses and Surveys, *Studies on Medical and Population Subjects No. 29. Cancer Mortality, England and Wales, 1911–70*, H.M.S.O., 1975.
2. Office of Population Censuses and Surveys, *Monitor D.H. 1 80/3. Cancer. Mortality in England and Wales, 1971–1978*, H.M.S.O., 1980.
3. Office of Population Censuses and Surveys, *Series DH2 no. 6. Mortality statistics. Cause*, H.M.S.O., London, 1980.
4. Office of Population Censuses and Surveys, *Series DH2 no. 7. Mortality statistics. Cause*, H.M.S.O., London, 1981.
5. Osmond, C., Gardner, M. J. and Acheson, E. D. 'An analysis of trends in cancer mortality in England and Wales during 1951–80 separating changes associated with period of birth and period of death', *British Medical Journal*, **284**, 1005–1008 (1982).
6. Osmond, C., Gardner, M. J., Acheson, E. D. and Adelstein, A. M. 'Trends in cancer mortality in England and Wales during 1951–80', *O.P.C.S. Series DH1* (in preparation).
7. Case, R. A. M. 'Cohort analysis of mortality rates as an historical or narrative technique,' *Brit. J. Prev. Soc. Med.*, **10**, 159–171 (1956).
8. Benjamin, B. 'Demographic aspects of ageing', in Yapp, W. B. and Bourne, G. H. (Eds), *The Biology of ageing* Symposia of the Institute of Biology, No. 6. The Institute of Biology, London 1957.
9. Nordling, C. O. 'A new theory of the cancer-inducing mechanism', *British Journal of Cancer*, **1**, 68–72 (1953).
10. Beral, V. 'Cancer of the cervix: a sexually transmitted infection?', *Lancet*, May 25, 1037–1039 (1974).
11. Walter, S. D., Miller, C. T. and Lee, J. A. H. 'The use of age-specific mean cohort slopes in the analysis of epidemiological incidence and mortality data', *J. R. Statist. Soc. A.*, **139**, part 2, 227–245 (1976).
12. Hanson, M. R., McKay, F. W. and Miller, R. W. 'Three-dimensional perspective of U.S. cancer mortality', *Lancet*, 246–248 (1980).
13. Barrett, J. C. 'Age, time and cohort factors in mortality from cancer of the cervix', *J. Hyg. Camb.* **71**, 253–259 (1973).
14. Kermack, W. O., McKendrick, A. G. and McKinlay, P. L. 'Death rates in Great Britain and Sweden: expression of specific mortality rates as products of two factors, and some consequences thereof', *Journal of Hygiene*, **34**, 433–457 (1934).
15. Spicer, C. C. 'The generation method of analysis applied to mortality from respiratory tuberculosis', *Journal of Hygiene*, **52**, 361–369 (1954).
16. Bjarnason, O., Day, N., Snaedel, G. and Tulinius, H. 'The effect of year of birth on the breast cancer age-incidence curve in Iceland', *Int. J. Cancer*, **13**, 689–696.
17. Breslow, N. E. and Day, N. E. 'Indirect standardisation and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data' *J. Chron. Dis.*, **28**, 289–303 (1975).

18. Barrett, J. C. 'The redundant factor method and bladder cancer mortality', *J. of Epidemiology and Community Health*, **32**, 314–316 (1978).
19. Barrett, J. C. 'A method of mortality analysis: application to breast cancer', *Rev. Epidem. et Sant Publ*, **26**, 419–425 (1978).
20. Barrett, J. C. 'Duration, period and cohort measures in marital fertility', *Population et Famille*, **47**, 69–79 (1979).
21. Barrett, J. C. 'Cohort mortality and prostate cancer', *J. Biosoc. Sci.*, **12**, 341–344 (1980).
22. Price, D. O. 'A respecification of variables in cohort analysis'. *Amer. Stat. Assoc. Meetings*, 23–26 Aug, 689–694 (1976).
23. Greenberg, B. G., Wright, J. J. and Sheps, C. G. 'A technique for analysing some factors affecting the incidence of syphilis.' *Journal of the American Stat. Assoc.*, **45**, 373–399 (1950).
24. Beard, R. E. 'Actuarial methods of mortality analysis', *Proceedings of the Royal Society of London, Series B*, **159**, 56–64 (1963).
25. Day, N. E. and Charnay, B. 'Time trends, cohort effects, and ageing as influence on cancer incidence', in Magnus, K. (ed.), '*Trends in Cancer Incidence, Causes and Practical Implications*', Hemisphere Publishing Corporation, New York, 1982.
26. Fienberg, S. E. and Mason, W. M. 'Identification and estimation of age-period-cohort models in the analysis of discrete archival data', University of Minnesota School of Statistics Technical Report No 286, 1977.
27. Sacher, G. A. 'Analysis of life tables with secular terms', in *The Biology of Ageing*, American Institute of Biological Sciences Symposium no. 6, 253 (1960).
28. Holman, C. D. J., James, I. R., Gattey, P. H., Armstrong, B. K. 'An analysis of trends in mortality from malignant melanoma of the skin in Australia', *International Journal of Cancer*, **26**, 703–709 (1980).
29. Pullum, T. W. 'Parametrizing age, period and cohort effects: An application to U.S. delinquency rates, 1964–73', in Schuessler, K. F. (Ed.) 'Sociological Methodology 1978' Jossey-Bass (San Francisco) 116–140 (1978).
30. Holman, C. D. J., James, I. R., Segal, M. R. and Armstrong, B. K. 'Recent trends in mortality from prostate cancer in male populations of Australia and England and Wales', *Br. J. Cancer*, **44**, 340–348 (1981).
31. James, I. R. and Segal, M. R. 'A method of mortality analysis incorporating age-year interaction, with application to prostate cancer', *Biometrics* (1982) (in press).
32. Mason, K. O., Mason, W. M., Winsborough, H. H. and Poole, W. K. 'Some methodological issues in cohort analysis of archival data', *Amer. Soc. Review.*, **38**, 242–258 (1973).
33. Glenn, N. D. 'Cohort analysts' futile quest: statistical attempts to separate age, period and cohort effects', *Amer. Soc. Review.*, **41**, 900–904 (1976).
34. Knoke, D. and Hout, M. 'Reply to Glenn', *Amer. Soc. Review.*, **41**, 905–908 (1976).
35. Mason, W. M., Mason K. O. and Winsborough, H. H. 'Reply to Glenn', *Amer. Soc. Review.*, **41**, 904–905 (1976).
36. Goldstein, H. 'Age, period and cohort effects—a confounded confusion', *Bulletin in Applied Statistics* **6**, 19–24 (1979).
37. Hobcraft, J. and Gilks, W. R. 'Age, period and cohort analysis effects in mortality studies', *International Union for Scientific Study of Population*, Meeting in Dakar, Senegal, 1981 (in press).
38. Case, R. A. M., Hosker, M. E., McDonald, D. B. and Pearson, J. T. 'Tumours of the urinary bladder in workmen engaged in the manufacture and use of certain dyestuff intermediates in the British Chemical Industry', *Brit. J. Industr. Med*, **11**, 75–104 (1954).
39. Lee, P. N. (ed). 'Statistics of smoking in the United Kingdom.' Research Paper I, 7th Edition, Tobacco Research Council, London, 1976.
40. Wald, N., Doll, R. and Copeland, G. 'Trends in tar, nicotine, and carbon monoxide yields of U.K. cigarettes manufactured since 1934', *Brit. Med. J*, **282**, 763–765 (1981).
41. Todd, G. F., Lee, P. N. and Wilson, M. J. 'Cohort analysis of cigarette smoking and of mortality from four associated diseases', *Occasional paper 3*, Tobacco Research Council, London, 1976.
42. Adelstein, A. M. 'Encouragement from recent statistics', in Raven, R. W. (ed). *'Outlook on Cancer'*, Plenum Press, New York and London, 1977.