STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:3018–3045
Published online 18 December 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2764

Age–period–cohort models for the Lexis diagram

B. Carstensen*, †

Steno Diabetes Center, Niels Steensens Vej 2, DK 2820 Gentofte, Denmark

SUMMARY

Analysis of rates from disease registers are often reported inadequately because of too coarse tabulation of data and because of confusion about the mechanics of the age–period–cohort model used for analysis. Rates should be considered as observations in a Lexis diagram, and tabulation a necessary reduction of data, which should be as small as possible, and age, period and cohort should be treated as continuous variables. Reporting should include the absolute level of the rates as part of the age-effects.

This paper gives a guide to analysis of rates from a Lexis diagram by the age–period–cohort model. Three aspects are considered separately: (1) tabulation of cases and person-years; (2) modelling of age, period and cohort effects; and (3) parametrization and reporting of the estimated effects. It is argued that most of the confusion in the literature comes from failure to make a clear distinction between these three aspects. A set of recommendations for the practitioner is given and a package for R that implements the recommendations is introduced. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: age–period–cohort model; Lexis diagram; follow-up studies; Poisson model; parametrization; epidemiology; demography

1. INTRODUCTION

Disease registries collect information on disease occurrence in populations by recording new cases by diagnosis, sex, age and date of diagnosis, etc. Description of disease rates by age and time is best conceptualized by regarding the observations in a Lexis-diagram.

The age–period–cohort model is a *descriptive* tool for observations from a Lexis diagram, typically from cancer registries or other disease registers. The model describes rates as a product of an age-effect, a period effect and a cohort effect. The aim is to give an overview of (1) the magnitude of the rates, (2) the variation by age and (3) time trends in the rates.

The database used as basis for the descriptive analysis is a tabulation of events (deaths, diagnoses of disease) and population size over a certain time period and age span, possibly restricted to certain birth cohorts.

*Correspondence to: B. Carstensen, Steno Diabetes Center, Niels Steensens Vej 2, DK 2820 Gentofte, Denmark.
†E-mail: bxc@steno.dk, URL: http://www.biostat.ku.dk/~bxc

Received 17 May 2006
Accepted 11 October 2006

The classical age–period–cohort model has been formulated as a model for this table where the effect of age, period and cohort are modelled as factors, i.e. with one parameter per level in the tabulation. In order to keep the number of parameters at a manageable level and to obtain reasonably smooth curves for the effects, the tabulation has usually been by 5-year age and period intervals.

Since the date of diagnosis is the sum of the date of birth and the age at diagnosis, there will be a constraint in any model which includes these three variables on a linear scale. The literature is abundant with attempts to overcome this so-called identifiability problem. It is of course futile to overcome the problem of having two variables as well as their sum in the same linear model. The identifiability problem is not much different from any other problem that arises from convenience formulation of models by over-parametrization as is, for example, often the case with the two-way ANOVA model. As such it cannot be solved properly without a view to the subject matter.

In this paper I formulate the age–period–cohort model as a general model for observation in the Lexis diagram. Clayton and Schifflers [1, 2] gave a careful exposition of the modelling problems in this setting, in particular, advice on what functions of the rates that could be (meaningfully) estimated. Keiding [3] gave an exposition of the analysis of rates in the Lexis diagram in continuous time, under various observation schemes.

In this paper I will focus on practical aspects of analysis of data where events are derived from disease registers and population risk time from census data, typically from national statistical offices.

The main points of the paper are (1) the tabulation of data should be part of the data analysis and as little information as possible should be thrown away by tabulation, (2) age, period and cohort should be regarded as continuous variables and (3) the absolute levels of the rates should be a part of the reporting. This requires a formulation that allows any kind of tabulation of data, not only by age and date of event, but also by date of birth, and not necessarily in intervals of the same length. I will separate the issues of the *tabulation* of the data, the *model* for the age, period and cohort effects and the *parametrization* of these. These three issues are mixed up in many papers discussing the models, mainly because the tabulation of data has been taken for granted, and the default model has been a factor model.

Section 2 gives a brief overview of the initial steps of an analysis of the observations in a Lexis diagram, Section 3 discusses tabulation of cases and person-years by age, period and cohort, and the implications for analysis. In Section 4 the Poisson model for rates is briefly reviewed and the options for modelling the *underlying* rates from the Lexis diagram discussed. The core of the paper is in Sections 5 and 6, where the parametrization of the age–period–cohort model is discussed *without* reference to tabulation of data and the particular parametric form chosen for the three effects. Finally, I give a few remarks about graphical display of results in Section 8 and in Section 9 I use data on testis cancer incidence in Denmark to illustrate the options given. In the discussion in Section 10, a set of recommendations are given for practical analysis and reporting of rates observed in a Lexis diagram. Two appendices address technical details for person-years calculation and matrix algebra for construction of parameter estimates.

2. DESCRIPTION OF RATES

2.1. Initial plots

Prior to analysis by age–period–cohort models one should always plot the observed rates. This will of course require that cases and person-years be tabulated in classes that are sufficiently large

DOI: 10.1002/sim

to produce fairly stable rates. There are four classical plots that should be made.

1. Rates *versus* age, observations within each period connected, i.e. cross-sectional age-specific rates.
2. Rates *versus* age, observations within each birth cohort connected, i.e. longitudinal age-specific rates.
3. Rates *versus* period, observations within each age-class connected.
4. Rates *versus* cohort, observations within each age-class connected.

These four plots are initial explorations of whether rates are proportional between periods or cohorts. If rates are plotted on a log scale the first and third will exhibit parallel lines if age-specific rates are proportional between periods (i.e. follow an age–period model), the second and fourth if they are proportional between cohorts (i.e. follow an age–cohort model).

These plots require a reasonably coarse tabulation to be informative. But the tabulation used for a first simple overview of data need not be the basis for the entire analysis.

An example of these four plots are given in Figure 1 for rates of testis cancer in Denmark. An important feature of these plots is the recognition of the *absolute* level of the rates.

The numbers used as basis for the plots are in Table I.

2.2. Modelling rates

There are three separate issues to consider when using a statistical model to describe rates from a disease register (observations in a Lexis diagram):

Data: How should data be tabulated: should we use 1-year intervals or 5-year intervals? Should the tabulation be by age and period, by age and cohort, by period and cohort or should it be by all three: age, period and cohort?

Model: How should we model the effects of age, period and cohort: use a factor model (one parameter per level) or smooth functions of the three variables? Should the smoothing be parametric or non-parametric?

Parametrization: How should we parametrize the estimated effects: what constraints should be used and how is it made clear which ones we have used? What is a sensible graphical display?

Note that I have separated the model and the parametrization of it. A (linear) model in my terminology is synonymous with the linear subspace spanned by the columns of the design matrix. Some authors (e.g. Clayton and Schifflers [1, 2]) have referred to different parametrizations of the same model as different models.

3. DATA

In principle the entire population could be regarded as a cohort, and data analysed in continuous time. However, this will rarely be feasible, so in practise data will be tabulated. The analysis data set will have one record per subset of the Lexis diagram, with number of events and risk time as outcome variables, and mean age, period and cohort as explanatory variables.

Any tabulation of data represents an information loss (rounding of age and date of diagnosis), so the tabulation should be as detailed as possible, it should only be limited by the availability of population figures. Cells with 0 events (cases) will not invalidate the analysis, so there is no lower limit to the tabulation intervals (except for computing capacity).

DOI: 10.1002/sim

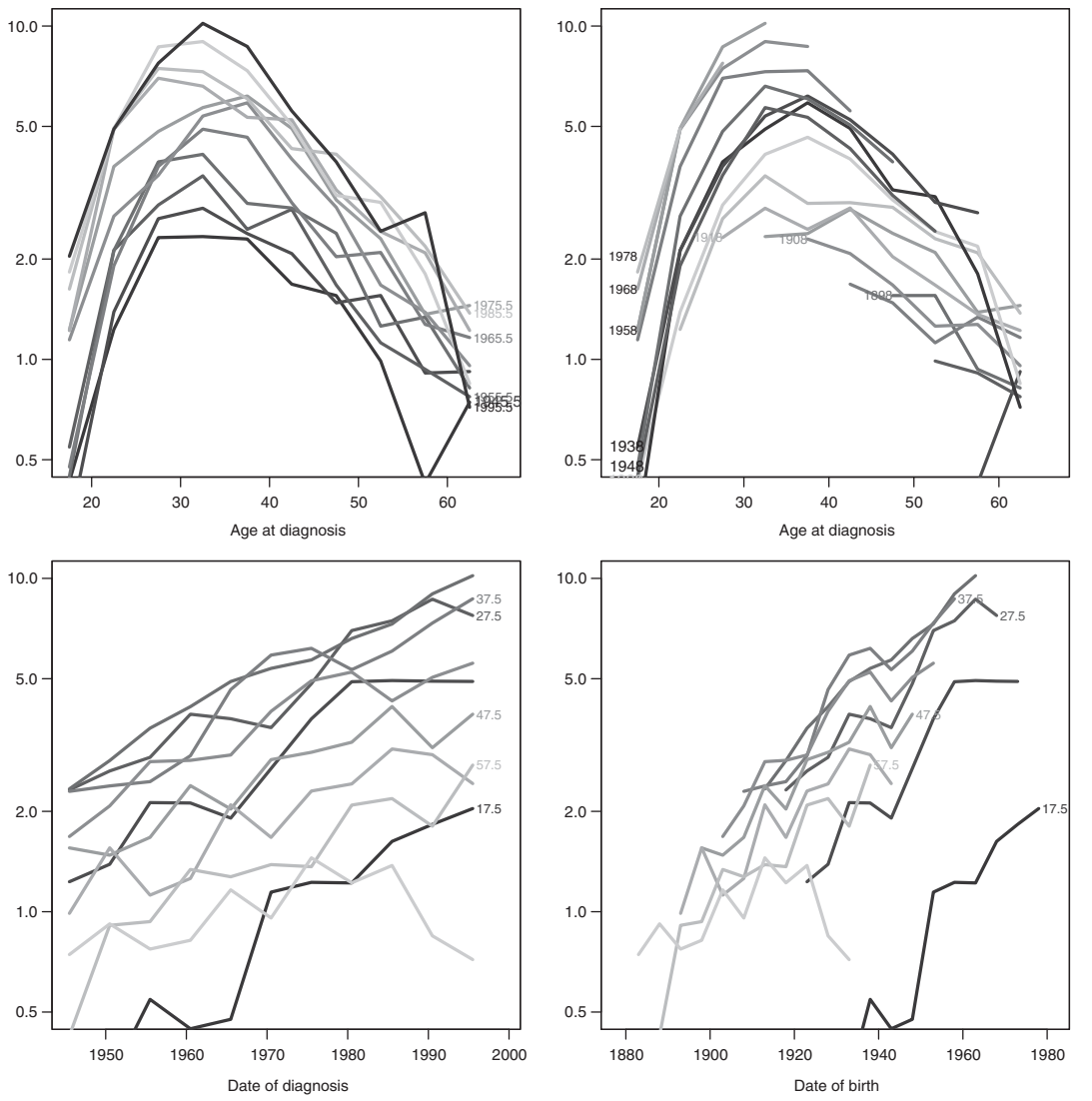

Figure 1. The classical four plots for rates observed in a Lexis diagram. Rates of testis cancer (per 100 000 person-years) in Denmark 1943–1997 in ages 15–64 years. Top left: age-specific rates by period of diagnosis. Top right: age-specific rates by date of birth. Bottom left: period-specific rates by age. Bottom right: cohort-specific rates by age.

3.1. Cases

If disease cases are taken from a register, cases can be tabulated arbitrarily by age at diagnosis, date of diagnosis (period) and date of birth (cohort). It will for example be possible to enumerate cases for each $1 \times 1 \times 1$-year triangle of the Lexis diagram as illustrated for Danish testis cancer cases in Figure 2. In principle, it would even be possible to produce a tabulation of data by single months as well.

DOI: 10.1002/sim

Table I. Cases of testis cancer and male person-years in Denmark 1943–1996 in 5-year classes.

	Mean period										
Mean age	1945.5	1950.5	1955.5	1960.5	1965.5	1970.5	1975.5	1980.5	1985.5	1990.5	1995.0
						Cases					
17.5	10	7	13	13	15	33	35	37	49	51	41
22.5	30	31	46	49	55	85	110	140	151	150	112
27.5	55	62	63	82	87	103	153	201	214	268	194
32.5	56	66	82	88	103	124	164	207	**209**	258	251
37.5	53	56	56	67	99	124	142	152	188	209	199
42.5	35	47	65	64	67	85	103	119	121	155	126
47.5	29	30	37	54	45	64	63	66	92	86	96
52.5	16	28	22	27	46	36	50	49	61	64	51
57.5	6	14	16	25	26	29	28	43	42	34	45
62.5	9	12	11	13	20	18	28	23	26	15	10
						Person-years (*in 1000s*)					
17.5	2321	2233	2382	2919	3155	2883	2858	3033	3015	2789	2011
22.5	2439	2234	2165	2313	2881	3162	2902	2859	3059	3052	2283
27.5	2372	2345	2169	2096	2294	2888	3168	2883	2869	3095	2507
32.5	2398	2324	2308	2135	2100	2310	2881	3136	**2865**	2871	2464
37.5	2308	2349	2281	2281	2135	2107	2302	2856	3107	2846	2292
42.5	2082	2263	2305	2250	2270	2129	2090	2273	2821	3071	2264
47.5	1866	2030	2214	2260	2214	2239	2095	2047	2229	2770	2453
52.5	1618	1801	1962	2146	2198	2155	2173	2027	1982	2163	2105
57.5	1413	1538	1713	1868	2042	2095	2051	2059	1923	1883	1634
62.5	1210	1305	1424	1584	1720	1880	1930	1884	1890	1772	1392

Note: The mean date in the period 1943–1947 is 1945.5, etc. The last period 1993–1996 is only 4 years, so the mean date is here 1995.0. The bold-face entry in the table is further subdivided in 50 subsets by one-year classes of age, period and cohort in Figure 1.

3.2. Person-years

Population figures are needed to produce rates, and the availability of these will normally be the limiting factor. In most countries, population figures in 1-year age classes for each calendar year are available. Such figures of population prevalence can be used to compute the risk time (person-years) in triangular subsets of the Lexis diagram.

For the sake of simplicity, the following formulae are given for one-year age classes and one-year periods.[‡] Specifically, let $\mathsf{L}_{a,p}$ be the population size in age a, at beginning of year p, see the left part of Figure 3, where $a = 61$ and $p = 1980$. The risk time in age class $a-1$, during year p, among those aged $a-1$ at the beginning of year p (i.e. born in year $p-a$, triangle A in Figure 3) is estimated as

$$y_{a-1,p,p-a} = \tfrac{1}{3}\mathsf{L}_{a-1,p} + \tfrac{1}{6}\mathsf{L}_{a,p+1}$$

[‡]Strictly speaking these formulae are wrong as they give *number* of cases rather than *risk time*. In order to make them correct all quantities must be multiplied by the interval length, in this case 1 year.

DOI: 10.1002/sim

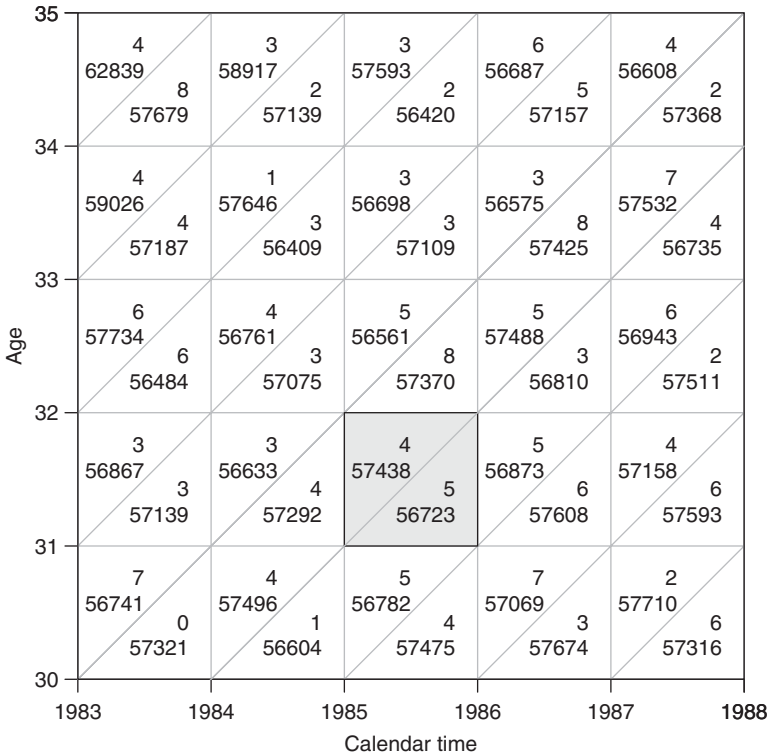

Figure 2. Danish testis cancer cases and male person-years in ages 30–34 and period 1983–1987, subdivided by age, period and cohort in 1-year classes (subdivision of the bold face entry in Table I). For example, in the highlighted square, there were 9 cases diagnosed in age 31 in the year 1985; 4 born in 1953 (upper triangle, persons who cross the 32 age line in 1985) and 5 born in 1954 (lower triangle, persons who cross the 31 age line in 1985).

and in the corresponding lower triangle (i.e. among those born in year $p-a$, triangle B in Figure 3):

$$y_{a,p,p-a} = \tfrac{1}{6}\mathsf{L}_{a-1,p} + \tfrac{1}{3}\mathsf{L}_{a,p+1}$$

These formulae are based on an assumption of constant mortality in each triangle of the Lexis diagram. The derivation of these formulae are to my knowledge first seen in a set of lecture notes from Oslo University by Sverdrup [4]. A precise derivation of the formulae is given in Appendix A.

The formulae should also be used for calculation of risk time in rectangles of the Lexis diagram as well. The total risk time in age-class a and period p is best estimated by

$$y_{a,p} = \tfrac{1}{6}\mathsf{L}_{a-1,p} + \tfrac{1}{3}\mathsf{L}_{a,p} + \tfrac{1}{3}\mathsf{L}_{a,p+1} + \tfrac{1}{6}\mathsf{L}_{a+1,p+1}$$

and not by $\frac{1}{2}\mathsf{L}_{a,p} + \frac{1}{2}\mathsf{L}_{a,p+1}$ as is commonly used. The two methods will give identical results if the mortality is 0 or birth rates are constant over time and mortality rates constant over age. Thus, the discrepancy will be largest in the older ages.

DOI: 10.1002/sim

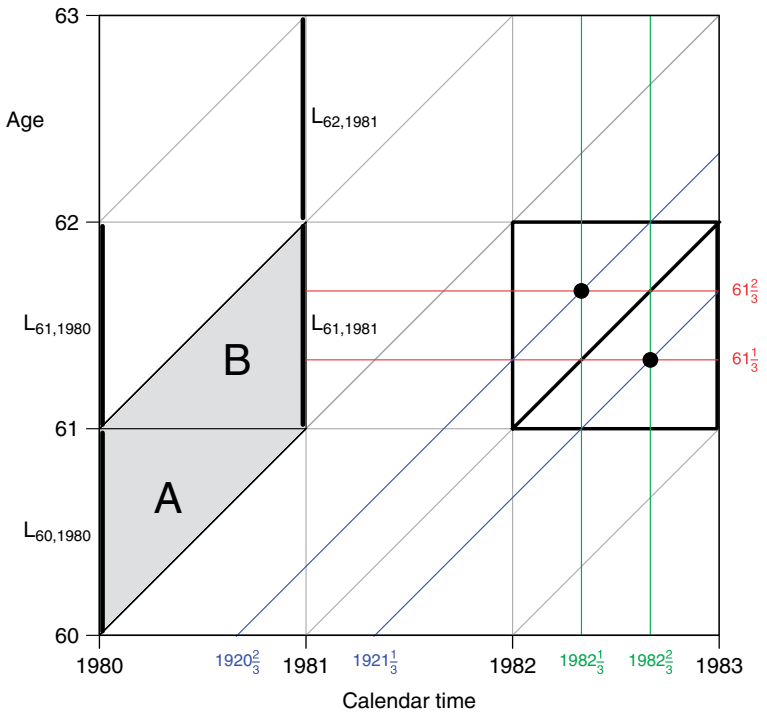

Figure 3. Lexis diagram. The thick lines in the left part show the population figures at the beginning of 1980 and 1981 necessary to estimate the population risk time in the triangles A and B. The right part of the diagram shows the mean age, period and cohort in the triangular subsets of a Lexis diagram. Note the connection between age, period and cohort: $p = c + a$: $1982\frac{1}{3} = 1920\frac{2}{3} + 61\frac{2}{3}$ and $1982\frac{2}{3} = 1921\frac{1}{3} + 61\frac{1}{3}$.

3.3. Mean age, period and cohort in triangular subsets

If a tabulation is by age and period (A-sets: □), by period and cohort (B-sets: ⌭) or by age and cohort (C-sets: ▱), the mean age, period and cohort in each set is simply the mean for the corresponding tabulation intervals for each of the two tabulation variables. The mean for the third variable is obtained using the relation $a = p - c$.

But when subdividing the Lexis diagram in triangles, the mean age, period and cohort in a given set is not equal to the mean of the classes chosen for tabulation. The means are offset by $\frac{1}{6}$ of the tabulation interval, as shown in the right part of Figure 3, see e.g. [5]; a formal derivation of this given in [6]. These values should be used when modelling rates based on a tabulation in triangles. Note in particular that the relationship $a = p - c$ must hold for all subsets of the Lexis diagram (and hence for all units in the data set).

In the following there will be no assumptions about the particular shape of the subsets of the Lexis diagram used for tabulation, neither w.r.t. the number of tabulation variables, whether the tabulation intervals are equally long for different variables nor whether all subsets have the same shape. Thus, we consider rates observed as (D, Y) in arbitrary subsets of a Lexis diagram, where D is number of cases and Y the amount of risk time. The associated covariates are mean age a, mean period p, and mean cohort $c = p - a$ for the subsets.

DOI: 10.1002/sim

4. MODELS

The general form of a multiplicative age–period–cohort model for rates $\lambda(a, p)$ at age a in period p for persons in birth cohort $c = p - a$ is

$$\log[\lambda(a, p)] = f(a) + g(p) + h(c) \tag{1}$$

Here it is assumed that a, p and c represent the *mean* age, period and cohort for the observational units. The model allows the effects of each of the three variables to be non-linear (on the log rate scale). The particular parametric form chosen for the functions f, g and h is immaterial at this point.

4.1. 'Poisson' model

For tabulated data, one must assume that the rate is constant within each tabulation category (subset of Lexis diagram). The log likelihood contribution from observation of the random quantity (D, Y) in one subset is

$$l(\lambda|D, Y) = D\log(\lambda) - \lambda Y$$

Except for a constant ($D\log(Y)$) not involving the rate parameter this is the same as the log likelihood for an observation of a random variable D from a Poisson distribution with mean λY. The log likelihood for the entire table of (D, Y) is the sum of such terms, because individuals are assumed to be independent, and the contributions to different cells from one individual are *conditionally* independent. Hence, models for λ can be fitted using a programme for Poisson regression for independent observations, that allows for an *offset* term to separate the person-years from the rate in the expression for the mean.

However, the fact that the Poisson model and the constant rate model has the same likelihood does not mean that the number of cases is Poisson distributed, and in particular not that the amount of risk time is fixed. The Poisson machinery should only be used for making likelihood based inference. Inference based on the *distributional* properties of the Poisson is not necessarily correct.

The information about $\theta = \log(\lambda)$ is computed as minus the second derivative of the log likelihood evaluated at the ML-estimate:

$$l(\theta|D, Y) = D\theta - \mathrm{e}^{\theta}Y, \quad l'_{\theta}(\theta|D, Y) = D - \mathrm{e}^{\theta}Y, \quad l''_{\theta}(\theta|D, Y) = -\mathrm{e}^{\theta}Y$$

so $I(\hat{\theta}) = \mathrm{e}^{\hat{\theta}}Y = \hat{\lambda}Y = D$. Note that this is the *observed* information *not* the Fisher information which is the *expected* information. In the Poisson model the expected information is λY (because Y is assumed known in the Poisson model), but in the constant rate model further assumptions about the observation plan is required to compute the expected value of D. A discussion of this in more depth is found in Keiding [3], who also gives references for the more probabilistic literature.

4.2. Submodels

One may argue that test of the terms in model (1) is irrelevant as it is a descriptive model. However, it has been customary to make formal tests of the effects.

The relevant submodels can conveniently be arranged in a sequence which gives all the relevant tests as comparisons between adjacent lines.

DOI: 10.1002/sim

Model	$\log[\lambda(a, p)]$
Age	$f(a)$
Age-drift	$f(a) + \delta c$
Age–cohort	$f(a) + h(c)$
Age–period–cohort	$f(a) + g(p) + h(c)$
Age–period	$f(a) + g(p)$
Age-drift	$f(a) + \delta p$

Note that the two drift models (i.e. with a log linear trend in period respective cohort) are identical: since $p = a + c$, a term δa can be separated from δp and absorbed into $f(a)$ giving the cohort drift model. Thus, the age-drift model is the intersection of the age–period and the age–cohort model, as pointed out by Clayton and Schifflers [1].

The deviance as output from most programmes can be used to derive the likelihood-ratio test for the model reductions. The deviance statistic is also commonly used in isolation to judge whether a particular model provides an adequate fit to the data. The deviance statistic is the likelihood-ratio test against the model with a completely freely varying interaction between age and period (or cohort). This is a meaningful test if the data represent a meaningful tabulation of the underlying process, but since any tabulation used for observations in a Lexis diagram is arbitrary so is the interaction model. Thus, the formal goodness of fit tests does not have an interpretation in terms of the original model for the rates as function of age, period and cohort in continuous time, and so is largely meaningless. If the analysis is based on coarsely tabulated data (5×5-year classes, say) it may be argued that the full model is the only natural extension of the age–period–cohort factor model, and thus provide a sensible test.

However, if data are tabulated in very small intervals the number of units in the analysis will increase to thousands and even if the modelling of age, period and cohort effects are fairly detailed, the degrees of freedom of the goodness of fit tests will increase dramatically, and thus more easily produce a significant statistic even if the average deviance contribution from each cell is more or less constant. Hence, the deviance statistic is a quantity depending on the chosen tabulation rather than on the adequacy of the model in describing the rates.

4.3. Classical approach to modelling effects

As an extreme way of accommodating the non-linearity of the effects of age, period and cohort, the usual approach to modelling effects uses one parameter per distinct value of a, p and c, by defining the variables as 'factors' (class variables).

As the tabulation of data becomes finer, the age–period–cohort modelling by one factor level for each distinct value of the three factors becomes unfeasible. This problem in age–period–cohort modelling emerges because the 'factor' approach insists that effects be modelled by one separate parameter for each distinct value of the tabulation factors age, period and cohort. The factor models are thus effectively models that let the tabulation induce the model.

The classical approach (which has largely emerged from cancer epidemiology) has been to define a tabulation sufficiently coarse to avoid an excess amount of parameters in the modelling; keeping the number of parameters of the models at a reasonable level has lead to a coarse tabulation of data, typically in 5-year intervals. Thus, there has been a feed-back loop between the tabulation of data and the modelling approach based on the concept of piecewise constant rates—the 'factor'-

DOI: 10.1002/sim

modelling approach. This may have been induced by limited availability of population figures or by limited computing capacity (initially the need to compute standardized rates by hand before the advent of proper modelling hard- and software).

4.3.1. Three-way tabulated data and the factor model. If data are tabulated by age, period and cohort, and the factor coding of effects is used, the model will fall into two disjoint parts, in the sense that the likelihood function will be a product of two terms, one only involving data from upper triangles ($\triangle$) and one only involving data from lower triangles ($\triangle$), with separate sets of parameters. This was pointed out by Osmond and Gardner [7], but to my knowledge no one has suggested a way to remedy this problem, other than pooling the triangular subsets to quadrangular. The following section provides a solution.

4.4. Smoothing with parametric functions

Since the three variables age, period and cohort are originally continuous variables it seems natural to model their effects by parametric smooth functions of the class means, for example:

- Splines, i.e. 1st, 2nd or 3rd degree polynomials in predefined intervals, constrained to have identical values and derivatives at interval boundaries (knots).
- Natural splines, 3rd degree splines constrained to be linear beyond the outermost knots.
- Fractional polynomials, combination of polynomials of various powers, including non-integer powers.

Any of these approaches gives columns of the design matrix for age, period and cohort effects, and as such are just (generalized) linear models.

If sufficient data are available there will be little difference between these approaches, the major question to address is the number of parameters to use for modelling each time-scale. Moreover, all the standard paraphernalia of penalizing the roughness of the effects is available for fine tuning the number of parameters and the location of knots. However, penalizing the roughness is not necessarily desirable in a descriptive demographic model where sudden changes in effects may be perfectly sensible, for example, due to changes in diagnostic practice.

Parametric smoothing avoids the problem of two separate models when using a factor model for data tabulated by age, period and cohort. Any factor model uses one parameter per cohort, even for the youngest and oldest where usually little data are present. With a parametric function of cohort it is possible to let the model reflect the information available in different cohorts.

If the number of parameters in the terms describing an effect equals the number of categories, then the model will be the same as the factor model, albeit parametrized differently. Hence, the parametric models are submodels of the classical 'factor' model.

Heuer [8] suggested to use restricted B-splines (natural splines) to model age, period and cohort effects in finely tabulated data. The point I suggest here is to use the class means directly as continuous variables to construct the splines, apparently similar to Holford's [9] approach. Essentially, that is what Heuer ends up with too, albeit only for rectangular subsets.

Heuer [8] suggested as a rule of thumb to use one knot per five years of the timescale. I find this suggestion too rigorous, certainly the number of events (= amount of information) must be relevant in deciding how many knots can be accommodated. Even if the cohort scale has a length

DOI: 10.1002/sim

which is the sum of the lengths for age and period, I would suggest that approximately the same number of knots be used for all three timescales. If there is a special interest in age-effects, say, the number of knots on this scale could be increased. Furthermore it should be considered to place the knots so that the (marginal) number of events is the same between them, rather than equidistantly, as the information is proportional to the number of events.

4.4.1. Non-parametric smoothers. If a non-parametric smoothing is used it is difficult to keep strict control over the parametrization. From the next sections it will be clear that access to the design matrix is essential for handling the parametrization of the effects of age, period and cohort. In particular, it is important to be able to access and manipulate the design matrix when predictions are made for rate-ratios relative to a specified point. Therefore, the non-parametric option is not considered further in this paper.

5. IDENTIFIABLE LINEAR TREND?

Holford [10] suggested extracting the linear trends from the age-, period- and cohort-parameters from a factor model by regressing age-parameters on age, period parameters on period and cohort parameters on cohort, and then report the residuals as age, period and cohort effects. This would give a display of the identifiable quantities on a recognizable scale. The three remaining parameters would then be the age-slope and the period/cohort slope and the intercept. The intercept would depend on the choice of reference point for the age and period or cohort. Holford also showed how this could be incorporated directly in the modelling by making a projection of the columns of parts of the design matrix.

Regression of the estimates for the age, period and cohort classes on age, period and cohort and extraction of the linear trend produces a set of parameters (functions) $\tilde{f}$, $\tilde{g}$ and $\tilde{h}$ that are 0 on average with 0 trend and are connected to the original parameters by

$$f(a) = \tilde{f}(a) + \mu_a + \delta_a a$$

$$g(p) = \tilde{g}(p) + \mu_p + \delta_p p$$

$$h(c) = \tilde{h}(c) + \mu_c + \delta_c c$$

Holford notes that $\delta_a + \delta_p$ and $\delta_p + \delta_c$ are invariants in the sense that regardless of the initial parametrization of f, g and h, they will have the same values, because any other parametrization $(\check{f}, \check{g}, \check{h})$ can be obtained by a suitable choice of μ_a, μ_c and γ as

$$\check{f}(a) = f(a) - \mu_a \qquad - \gamma a$$

$$\check{g}(p) = g(p) + \mu_a + \mu_c + \gamma p$$

$$\check{h}(c) = h(c) \qquad - \mu_c - \gamma c$$

It is easily verified that $\check{f}(a)+\check{g}(p)+\check{h}(p) = f(a)+g(p)+h(c)$ for any value of $(a, p, c = p-a)$, and that the regression slopes of $\check{f}$, $\check{g}$ and $\check{h}$ differ from those for f, g and h with the same numerical quantity (γ), but that the quantities $\delta_a + \delta_p$ and $\delta_p + \delta_c$ are the same.

DOI: 10.1002/sim

As Holford showed, the 'detrended' period estimates can be obtained directly in the model fitting by replacing the part of the design matrix corresponding to period by a matrix with columns orthogonal to the intercept column and the (period) drift column. This matrix is found by taking the original columns and projecting them on the orthogonal complement to the space spanned by the constant and the drift.

However, the uniqueness of the overall secular trend $\delta_p + \delta_c$ depends on the definition of 'orthogonal' or '0 on average with 0 slope'. The formulae devised by Holford are based on orthogonality with respect to the usual inner product

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum x_i y_i$$

However, this is not the only way of defining the drift, instead we could use an inner product of the type

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum x_i w_i y_i$$

with some pre-defined weights. One obvious choice would be to take the w_is proportional to the number of cases in each record (tabulation cell), i.e. using the observed information about the log rate as weights. One might argue that this choice depends on data and as such would render comparisons of trends across populations or regions invalid. However, the choice of the tabulation unit for the weighting is equally arbitrary; recall that the basic data unit after all is the follow-up of single persons in the population. With this in mind a choice would obviously be to choose the person-years as weights for the inner product. In most data sets these are options that put more or less weight in the older age-classes.

Hence, the linear components ($\delta_a + \delta_p$ and $\delta_p + \delta_c$) devised by Holford are just as arbitrary as any other set of extracted linear trends. The size of extracted linear trends are not a feature of the *model*; they are a feature of the model *and* the (arbitrarily) chosen method for extracting the linear trends, hence it is largely a matter of taste which inner product to use when extracting the trends.

6. PARAMETRIZATION OF THE AGE–PERIOD–COHORT MODEL

Since the aim is to model rates as a function of age, period and cohort, the logical approach to parametrization is to formulate the problem in continuous time, i.e. by a model for the rate at any point (a, p) in the Lexis diagram. A general formulation of the model is

$$\log[\lambda(a, p)] = f(a) + g(p) + h(c) \tag{2}$$

for three functions, f, g and h. The model predicts the rates at any point in the Lexis diagram. Applied to tabulated data the rates are assumed to be constant in each of the subsets in the Lexis diagram. The general form of the model makes parametric assumptions about the rates in these subsets. Note that the classical factor approach to modelling also falls under this formulation—the functions are just assumed piecewise constant in larger intervals.

The challenge is to choose a parametrization of these three functions in a way that:

1. is meaningful,
2. is understandable and recognizable,

DOI: 10.1002/sim

3. is practically estimable by standard software,
4. allows reconstruction of the fitted rates from the reported values.

Little attention has been paid to the last point, for example both Heuer [8] and Holford [9], present graphs for relative effects without considering the absolute level of the rates.

As noted by Holford [10] and later by Clayton and Schifflers [2], the only components of the model (2) that can be uniquely determined are the second derivatives of the three functions, and yet the relevant representation of the model is by graphs of three functions f, g and h that sum to the predicted log rates. The first derivatives as well as the absolute levels can be moved around between the functions. One choice is only to show the second derivatives, but as the scale for these is not easily understandable this is not an option in practise, although it has been used [11].

6.1. Choice of parametrization

For the sake of the argument, consider first the age–cohort model

$$\log[\lambda(a,c)] = f(a) + h(c)$$

In this model only the *first* derivatives (contrasts) of f and h are identifiable. This is traditionally fixed by choosing a reference cohort c_0, say, and constrain $h(c_0) = 0$. This will make $f(a)$ interpretable as the age-specific log rates in cohort c_0 and $h(c)$ as the log rate ratio of cohort c compared to cohort c_0.

The formalism behind this is to write

$$\log[\lambda(a,c)] = \tilde{f}(a) + \tilde{h}(c) = (f(a) + \mu) + (h(c) - \mu)$$

and by choosing $\mu = h(c_0)$ we get the desired functions as

$$\tilde{f}(a) = f(a) - h(c_0), \quad \tilde{h}(c) = h(c) - h(c_0)$$

which indeed has the property that $\tilde{f}(a) + \tilde{h}(c) = f(a) + h(c)$ and $\tilde{h}(c_0) = 0$. In practise this can be implemented by choosing the parametrization of the model carefully. In the case of a factor model for the effect of cohort, this is known as choosing the reference cohort to be c_0. This is a standard procedure when fitting linear models, but it is rarely recognized as the solution to an identifiability problem.

A similar machinery can be invoked to explicitly move the three unidentifiables in an age–period–cohort model around between f, g and h by choosing μ_a, μ_c and δ so that the resulting functions, $\tilde{f}(a)$, $\tilde{g}(p)$ and $\tilde{h}(c)$ meet some desired constraints:

$$\begin{aligned}\log[\lambda(a,p)] = \tilde{f}(a) + \tilde{g}(p) + \tilde{h}(c) = f(a) &- \mu_a &&- \gamma a\\ &+ g(p) &+ \mu_a &+ \mu_c + \gamma p\\ &+ h(c) &&- \mu_c - \gamma c\end{aligned}$$

In the age–period–cohort model with three terms and where only the *second* derivatives of the effects are identifiable, two levels and one value of the *first* derivative must be fixed. This is frequently done by fixing one value on the cohort scale to be 0 and two points on the period scale

DOI: 10.1002/sim

to be 0, thereby fixing the overall slope of the period parameters, but a less technical principle for the choice of parametrization is desirable.

6.2. A principle for parametrization

Any parametrization of the age–period–cohort model fixes two levels and a slope among the three functions, but different principles can be invoked to accomplish this.

One principle for choice of parametrization is based on an extension of the assumptions behind way the age–cohort model was parametrized:

1. The age-function should be interpretable as log age-specific rates in cohort c_0 after adjustment for the period effect.
2. The cohort function is 0 at a reference cohort c_0, interpretable as log RR relative to cohort c_0.
3. The period function is 0 on average with 0 slope, interpretable as log RR relative to the age–cohort prediction (residual log RR).

   Alternatively, the period function could be constrained to be 0 at a reference date, p_0. In this case the age-effects at $a_0 = p_0 - c_0$ would equal the fitted rate for period p_0 (and cohort c_0), and the period effects would be residual log RRs relative to p_0.

The first choice fixes one constant (0 at c_0), and the third fixes a level (0 on average or 0 at p_0) and a slope (0 slope for the period function). The inclusion of the slope (drift) with the cohort effect makes the age-effects interpretable as cohort-specific rates of disease (longitudinal rates). Depending on the subject matter, the role of cohort and period could be interchanged, in which case the age-effects would be cross-sectional rates for the reference period. Heuer [8] also discuss these two interpretations of the age-effects depending on whether the drift is allocated with period or cohort.

In practise this can be implemented as follows. The point is to choose the columns of the design matrix in such a way that desired functions are simple linear combinations of the estimated parameters, which will give a simple calculation of standard errors and hence confidence intervals. This is detailed in Section 6.3.

Suppose functions $\hat{f}$, $\hat{g}$ and $\hat{h}$ have been estimated from data. Then we fix a reference cohort c_0, and extract the linear part from $\hat{g}$, for example, by regressing the values of $\hat{g}(p)$ on p

$$\tilde{g}(p) = \hat{g}(p) - (\mu + \beta p)$$

where μ is chosen to make $\tilde{g}(p)$ equal to 0 on average. The log rates can then be expressed in three new terms

$$\log[\lambda(a, p)] = \tilde{f}(a) + \tilde{g}(p) + \tilde{h}(c)$$

where

$$\begin{aligned}
\tilde{f}(a) &= \hat{f}(a) + \mu + \beta a + \hat{h}(c_0) + \beta c_0 \\
\tilde{g}(p) &= \hat{g}(p) - \mu - \beta p \\
\tilde{h}(c) &= \hat{h}(c) + \beta c - \hat{h}(c_0) - \beta c_0
\end{aligned}$$

The functions $\tilde{f}(a)$, $\tilde{g}(p)$ and $\tilde{h}(c)$ defined this way fulfils the requirements above; $\tilde{g}(p)$ is 0 on average with 0 overall slope, $\tilde{h}(c)$ is 0 for $c = c_0$, and $\tilde{f}(a)$ has the dimension of log rates,

DOI: 10.1002/sim

referring to cohort c_0. If we prefer to fix the period function to 0 at a given date p_0, we just use $\mu = \hat{g}(p_0) - \beta p_0$, instead of using the estimated μ from the regression.

Note that in the derivation above, we made no assumptions about the algorithm used to extract the linear trends. It was only assumed that it was possible to de-trend g and h in some way.

6.2.1. Explicit drift parameter. A variant of the above approach is to extract the drift entirely and report it as a separate parameter and then report both cohort and period effects as 'residuals'. This would correspond to the partitioning of the model into terms:

$$\begin{aligned}\log(\lambda(a, p)) &= \tilde{f}_c(a) + \delta(c - c_0) + \tilde{g}(p) + \tilde{h}(c)\\ &= \tilde{f}_p(a) + \delta(p - p_0) + \tilde{g}(p) + \tilde{h}(c)\end{aligned}$$

where $\tilde{g}(p)$ and $\tilde{h}(c)$ are 'de-trended', i.e. have 0 slope. $\tilde{f}_c(a)$ are the age-specific rates in the reference cohort c_0 and $f_p(a)$ the age-specific rates in the reference period p_0. Thus, age-specific rates can be chosen to refer to either a specific cohort (longitudinal rates) or a specific period (cross-sectional rates). Note that $\tilde{f}_c(a) = \tilde{f}_p(a) + \delta(a - (p_0 - c_0))$, so if there is a positive drift ($\delta > 0$) the cohort (longitudinal) age-curve will be steeper than the period (cross-sectional) age curve.

6.3. Parametric models in practise

The goal of the parametrization is to produce estimates of three functions showing the age, the period and the cohort effects constrained in a sensible way. We also want confidence intervals for these. This section details how the design matrix for the model can be set up in such a way that derivation of these functions is done by simple linear functions of three separate subsets of parameters.

Consider first a simple age–cohort model with say cohort c_0 as reference. If a factor model is used we would set up a design matrix with one indicator column per age class and one indicator column for each cohort *except* for c_0. The estimated parameters would then be the log rates in each age class for cohort c_0, and log RRs relative to this.

If we instead model the age-specific log rates as quadratic in age we replace the age-columns with the three columns $[1|a|a^2]$ where a is the midpoint in each age-category. If the estimates for these three columns are $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)$, then the estimated log rate for age a in cohort c_0, is $\hat{\alpha}_0 + \hat{\alpha}_1 a + \hat{\alpha}_2 a^2$, or in matrix notation

$$(1 \;\; a \;\; a^2) \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$$

If the estimated variance–covariance matrix of $(\alpha_0, \alpha_1, \alpha_2)$ is $\mathbf{\Sigma}$ (a 3×3 matrix), then the variance of the log rate at age a is

$$(1 \;\; a \;\; a^2)\mathbf{\Sigma} \begin{pmatrix} 1 \\ a \\ a^2 \end{pmatrix}$$

DOI: 10.1002/sim

This rather tedious approach is an advantage if we simultaneously want to compute the estimated rates at several ages $a_1, a_2, \ldots, a_n$. The estimates and the variance covariance of these are then

$$\begin{pmatrix} 1 & a_1 & a_1^2 \\ 1 & a_2 & a_2^2 \\ \vdots & \vdots & \vdots \\ 1 & a_n & a_n^2 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & a_1 & a_1^2 \\ 1 & a_2 & a_2^2 \\ \vdots & \vdots & \vdots \\ 1 & \cdots & a_n^2 \end{pmatrix} \boldsymbol{\Sigma} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ a_1 & a_2 & \cdots & a_3 \\ a_1^2 & a_2^2 & \cdots & a_3^2 \end{pmatrix}$$

The matrix we use to multiply with the parameter estimates is the age-part of the design matrix we would have if observations were for ages $a_1, a_2, \ldots, a_n$. The product of this piece of the design matrix and the parameter vector represents the function $f(a)$ evaluated in the ages $a_1, a_2, \ldots, a_n$.

If a spline model for the age effect is used, the age-part of the design matrix will be columns of base vectors for the splines. If cubic splines are used and knots k_1 and k_2 are used the columns of the design matrix corresponding to age will be:

$$1, \quad a, \quad a^2, \quad a^3, \quad (a-k_1)_+^3, \quad (a-k_2)_+^3$$

with the notation $x_+ = \max(0, x)$. These functions of $a_1, a_2, \ldots, a_n$ will be multiplied with the six parameters $\alpha_0, \alpha_1, \ldots, \alpha_5$ in the spline model to give estimates of the log rates as $\hat{\alpha}_0 + \hat{\alpha}_1 a + \hat{\alpha}_2 a^2 + \hat{\alpha}_3 a^3 + \hat{\alpha}_4 (a-k_1)_+^3 + \hat{\alpha}_5 (a-k_2)_+^3$.

Now suppose that the cohort effect is modelled by a cubic spline as well, i.e. we use the columns

$$c \quad c^2 \quad c^3 \quad (c-k_1)_+^3 \quad (c-k_2)_+^3$$

This will not assure that the estimated effect of cohort is 0 for cohort c_0; if the parameter estimates for the splines for cohort effect are $\hat{\gamma}_0, \hat{\gamma}_1, \ldots$ the estimated cohort effect at c_0 will be

$$\hat{\gamma}_0 + \hat{\gamma}_1 c_0 + \hat{\gamma}_2 c_0^2 + \hat{\gamma}_3 c_0^3 + \hat{\gamma}_4 (c_0 - k_1)_+^3 + \hat{\gamma}_5 (c_0 - k_2)_+^3$$

Therefore, if we replace the cohort-columns in the design matrix with the columns

$$c - c_0, \quad c^2 - c_0^2, \quad c^3 - c_0^3, \quad (c-k_1)_+^3 - (c_0-k_1)_+^3, \quad (c-k_2)_+^3 - (c_0-k_2)_+^3$$

we get an estimated cohort effect which is 0 at c_0. What we have done is just to subtract a different constant from each column, which does not influence the model, it only changes the intercept parameter, which is a part of the age-effects. This way we automatically also get the age-parameters to refer to the log rates for cohort c_0.

A similar approach to parametrization fulfilling the requirements above can be implemented for the age–period–cohort model as follows. The idea is that we want to end up with three sets of columns that directly allows to compute age, period and cohort effects at any set of points we wish.

1. Set up model matrices for age, period and cohort, $\mathbf{M}_a$, $\mathbf{M}_p$ and $\mathbf{M}_c$, each including the intercept term. If a factor model is used these are just columns of indicators, one for each level of age, period and cohort. If a spline model is used the matrices will be columns of base vectors for the splines, as in the example above, or generated as a B-spline basis (see e.g. [8]).
2. Extract the linear trend from $\mathbf{M}_p$ and $\mathbf{M}_c$, by projecting their columns onto the orthogonal complement of $[1|p]$ and $[1|c]$, respectively. Estimates for period and cohort effects derived

DOI: 10.1002/sim

using these matrices will be 0 on average and with 0 overall slope. This projection is a non-trivial matrix operation, that is further detailed in Appendix II.

The resulting matrices $\tilde{\mathbf{M}}_p$ and $\tilde{\mathbf{M}}_c$ have two fewer columns than the original $\mathbf{M}_p$ and $\mathbf{M}_c$. This was in essence the proposal that Holford [10] made, also providing the projection matrix.

3. Centre the cohort effect around c_0: First take a row from $\tilde{\mathbf{M}}_c$ corresponding to c_0. Form a matrix of the same dimension as $\tilde{\mathbf{M}}_c$ with all rows equal to this, and subtract it from $\tilde{\mathbf{M}}_c$ to form $\tilde{\mathbf{M}}_{c_0}$.
4. The desired parametrization can now be obtained by using the $\mathbf{M}_a$ for the age-effects, $\tilde{\mathbf{M}}_p$ for the period effects and $[c - c_0|\tilde{\mathbf{M}}_{c_0}]$ for the cohort effects.

   Since the intercept is assumed to be included in $\mathbf{M}_a$, the age-effects are automatically adjusted for the centring of the cohort-effect around c_0, so they represent log rates for the cohort c_0.
5. The extracted drift is estimated separately by considering the estimated coefficient to the column $c - c_0$, and the standard error of it. If the drift is taken out as a separate parameter, both cohort and period effects will be residual effects constrained to be 0 on average with 0 slope.
6. Suppose the subsets of the estimated parameter vector corresponding to $\mathbf{M}_a$, $\tilde{\mathbf{M}}_p$ and $[c - c_0|\tilde{\mathbf{M}}_{c_0}]$ are $\hat{\alpha}_a$, $\hat{\beta}_p$ and $\hat{\gamma}_c$. The value of $\hat{f}(a)$ at the as actually present in the data set is $\mathbf{M}_a\hat{\beta}_a$. The variance of it is found by $\mathbf{M}_a^{\mathsf{T}}\hat{\mathbf{\Sigma}}_a\mathbf{M}_a$, where $\hat{\mathbf{\Sigma}}_a$ is the variance–covariance matrix of $\hat{\beta}_a$. In practical situations one would shave $\mathbf{M}_a$ down to a matrix with unique rows, each representing an observed point on the age-scale.

   The same machinery is used to derive effects for the other two effects at the observed points.

It is clear from the above that it will be convenient to have tools that can generate model matrices for the type of model used, as well as have facilities for matrix operations. These tools are available in Stata and SAS, but not directly integrated in the language as is the case with S-plus and R, where matrices may be entered directly in the fitting functions. A publicly available implementation of the methods given here is available in the function `apc.fit` in the `Epi` package for R. The `Epi` package also contain functions for matrix projection.

7. FITTING MODELS SEQUENTIALLY

It is possible to obtain an approximation to the parametrization outlined above using a small trick: first fit the age–cohort model. By omitting an explicit intercept and choosing a suitable reference for the cohort, the age-effect will be log rates for the reference cohort and the cohort effect will be log RRs relative to this. $\hat{f}(a)$ and $\hat{h}(c)$ are then used as age and cohort effects.

The log of the fitted values from this model is then used as offset variable in a model with period-effect

$$\log[\lambda(a, p)] = [\hat{f}(a) + \hat{h}(c)] + g(p)$$

The period effects from this model (also omitting an explicit intercept) are then used as the residual log RRs by period.

The estimates obtained by this sequential procedure are not the ML-estimates from the age–period–cohort model, they are marginal age–cohort estimates and period estimates *conditional*

DOI: 10.1002/sim

on the estimates from the age–cohort model, but in practise they will be very similar to the ML-estimates.

If one has an *a priori* assumption that mainly cohort-effects drive the change in rates, then this would the best way to model the rates, because the period effects would then only be residuals *conditional* on the estimated age and cohort effects.

Usually, the estimates from this approach will be close to the ML estimates, and the advantage is that the parametrization is very simple, no special manipulations are required to reparametrize and obtain standard errors. There are obvious extensions of this trick: first fit the age-drift model, and then sequentially cohort and period as 'residual' effects.

A variant of this procedure has been used by some authors as a way of fixing the age-effects to obtain identifiability [12]. The procedure is also implemented in the function `apc.fit` in the `Epi` package for R too.

8. GRAPHICAL DISPLAY OF EFFECTS

For any chosen constraints there will be three estimated functions, $\hat{f}(a)$, $\hat{g}(p)$ and $\hat{h}(c)$ which sum to the fitted log rates. These effects should be shown in one figure, with same equidistance

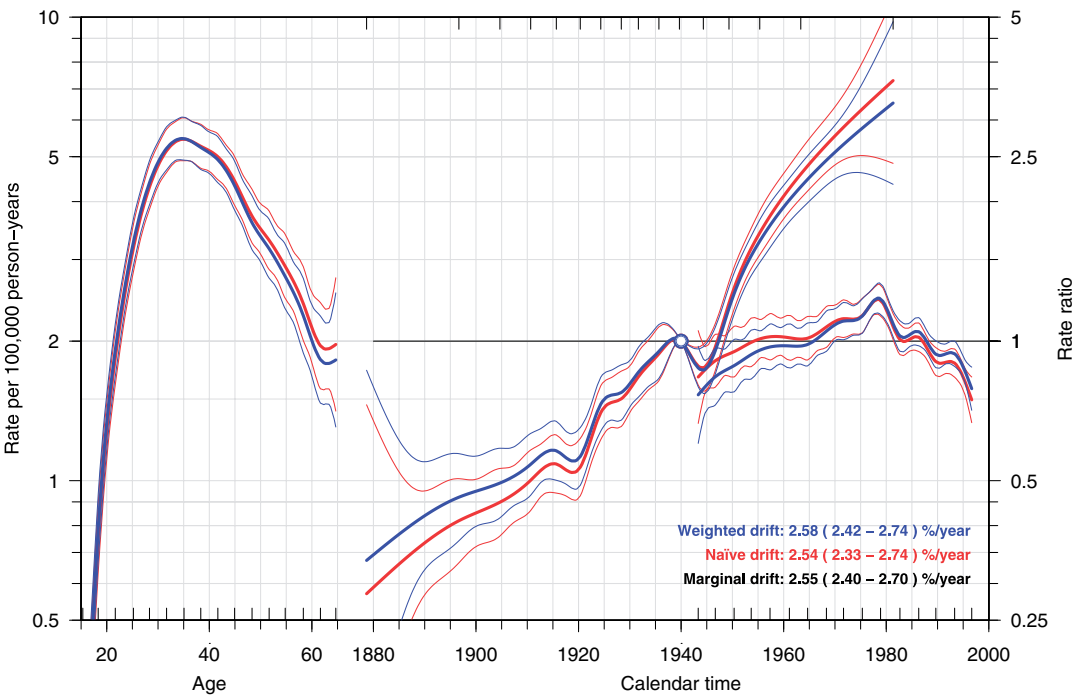

Figure 4. The estimated effects using the weighted and the naive approach to extracting the drift. For both approaches the drift is included with the cohort effect. The fit to the data is the same. The marginal drift is the drift estimate from the age-drift model. The inside tickmarks at the bottom and top indicate the placement of the knots for the natural splines. The cohort curve is the leftmost and longest of the two curves on the calendar time scale, the period curve the rightmost and shorter.

DOI: 10.1002/sim

for the horizontal scale for age, period and cohort. Also the relative extent for the rate-scale for age-effects and relative risk scale for period and cohort effects should be the same. This will put all three effects on a directly comparable scale and allow the slopes of the effects to be compared.

The figure must have the horizontal scale divided in two; one for age and one for cohort/period; the latter two will often cover overlapping calendar periods. The vertical scale will be a rate scale for the age-effect and a relative risk scale for the period and cohort effects. If a reference cohort or period is chosen a dot should be placed at $(c_0, 1)$ or at $(p_0, 1)$ to indicate this.

This is shown in Figure 4 for the Danish testis cancer data.

9. ANALYSIS OF DANISH TESTIS CANCER DATA

An annotated R-file and the Danish testis cancer data are available on my homepage `http://www.biostat.ku.dk/~bxc/APC/SiM-ex`.

The classical displays shown in Figure 1 produced with the function `rateplot`, which is the natural first step of the analysis, in this case based on data in 5 by 5 year classes. There is a clear tendency that cohort born around 1940–1945 show lower rates than those born earlier and later.

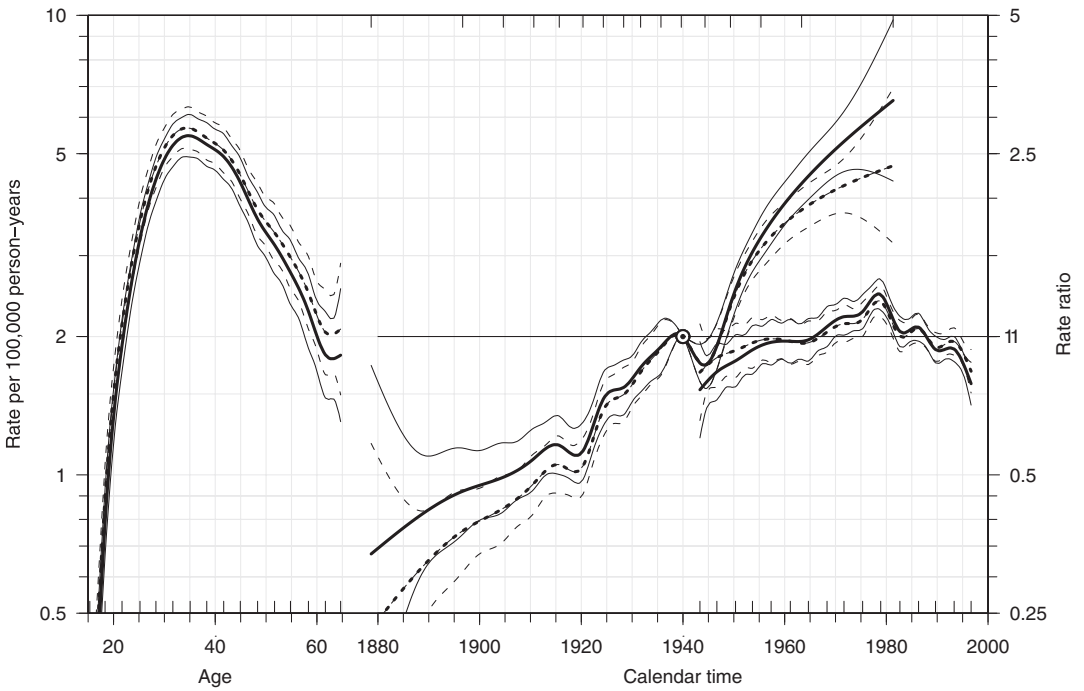

Figure 5. The estimated effects using the weighted approach to extracting the drift (full lines), contrasted with the sequential approach by first fitting the age–cohort model and then the period model to the residuals (broken lines). The age-effect refers to the 1940 cohort. The fit to the data is the not same for the two sets of estimates. In the sequential approach the age and cohort effects are from the age–cohort model, and some confounding of the cohort effect by period seems to be present. The cohort curve is the leftmost and longest of the two curves on the calendar time scale, the period curve the rightmost and shorter.

DOI: 10.1002/sim

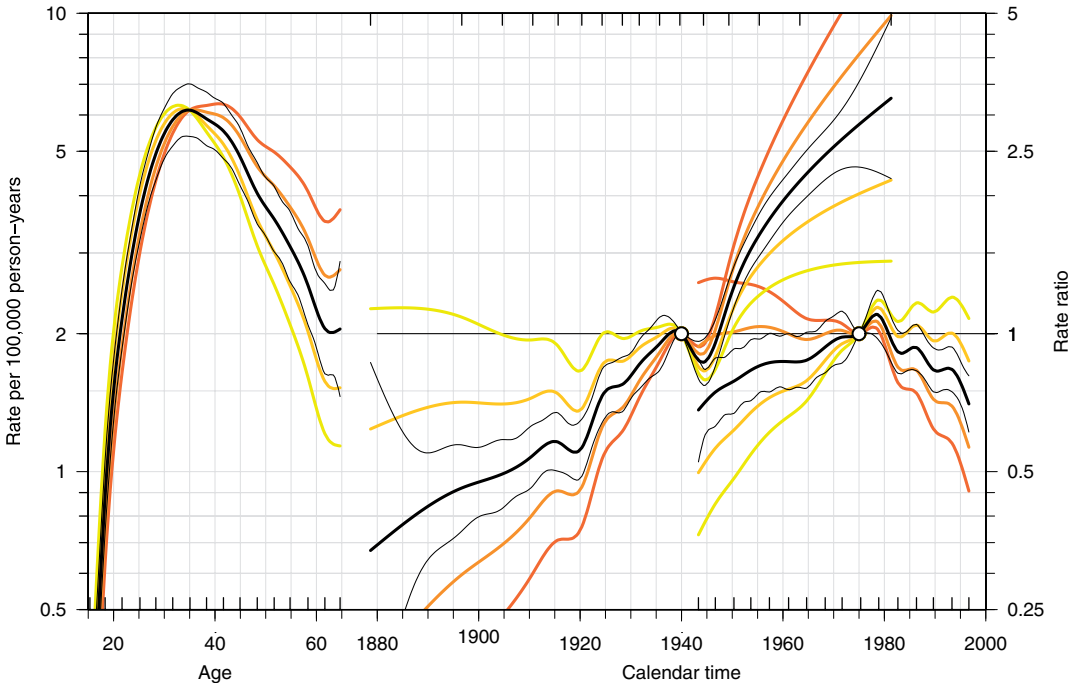

Figure 6. The estimated effects using the weighted approach to extracting the drift (black) and allocating it with the cohort, and using the 1940 cohort and 1975 period as references. Curves with added annual period drifts of $-2, -1, 1, 2\%$ are shown as well. The rates predicted from curves of like colours are the same. The cohort curve is the leftmost and longest of the two curves on the calendar time scale, the period curve the rightmost and shorter. A film-like version can be found at `www.biostat.ku.dk/~bxc/APC/Testis-film.pdf`.

Testis cancer cases in Denmark 1943–1996 were tabulated in 1-year classes by age, period and cohort. Population figures in 1-year age-classes at 1 January each year were obtained from Statistics Denmark. The risk time was computed as outlined in Section 3. The analysis is restricted to the age-classes 15–64 years.

The age–period–cohort model was fitted with `apc.fit`, which allows various models (linear splines, cubic splines, factors) and parametrizations to be used. The displays in Figures 4–6 shows a model where natural splines with 15 parameters for each of the effects were used, these were plotted with the functions `apc.frame` and `apc.lines`.

From the curves it is clear that there is a 'dip' in rates for the birth cohorts born during the first and second world wars. Such a dip is a second order feature of the curve and is therefore not an artefact of the parametrization. These two dips in the cohort effects are brought out clearly by a combination of the detailed tabulation of data, and the detailed parametrization of the cohort effect. Modelling the effects with fewer parameters would overlook the dip around the first world war.

10. DISCUSSION

Age–period–cohort models are descriptive tools for rates observed in a Lexis diagram. Proper analysis of rates should use the maximally available information. Therefore, tabulation in very

DOI: 10.1002/sim

coarse groups of age, period and cohort should be avoided. Whenever possible tabulation by all three variables should be done.

Overly coarse tabulation of data is abundant in the epidemiological literature, rates of childhood diabetes (ages 0–14 years) is for example commonly modelled using three 5-year age classes! [13–16].

At least for countries in western Europe, data on population size in 1-year age-classes will usually be available. The SEER programme in U.S.A. have made population data in one-year classes available at state level (see `http://seer.cancer.gov/popdata/`).

Large parts of the literature on the age–period–cohort models is difficult to read because of overly complicated notation, e.g. use of indexing of age, period and cohort groups by i, j and k running from 1 to I, J and K, giving rise to complicated indexing formulae, involving the total number of categories that are otherwise only relevant when specifying computer code. It seems more straightforward to use mnemonics like a, p and c and letting the indices be the mean of these continuous variable in each cell of the tabulation, even if takes much of the magic out of the subject. The use of the notational tradition from abstract mathematics has presumably distracted many readers from realizing that age–period–cohort analysis is about having two timescales (age and period) and their difference in the same model. Certainly, some authors seem to have been misled in this aspect, see e.g. [17, 18].

Effectively, only the factor model induced by the tabulation of data have been used in applied age–period–cohort modelling and it is also the predominant model addressed in the theoretical papers. Therefore, the parametrization problems have mostly been addressed in the framework of the factor model. This has led to a plethora of suggestions for parametrizations with little view to the principal aspects of the subject matter, see e.g. [19].

Heuer [8] suggested using splines in modelling the effects for rectangular subsets of the Lexis diagram, and although his formulation is tantamount to the use of the mean age, period and cohort for the subsets of the Lexis diagram, this point was not used in his exposition. Heuer also proposed the direct use of a projection (albeit only with respect to the common inner product) to generalize Holford's method to spline models; Holford [9] used spline modelling for rectangular subsets of the Lexis diagram too, but none of these authors included the rate dimension in the reporting of the models.

The tools available for age–period–cohort modelling have by and large been Poisson-modelling by programmes that produce a standard parametrization of factors such as `proc genmod` from SAS or the `glm` command from Stata. In most applied papers the authors have shied away from graphical reporting of the estimates [14, 15, 20]. This may be an indication of the technical problems associated with transforming the default parametrization from the statistical packages to a useful parametrization. Particularly, the derivation of standard errors of estimated curves can be a complicated task with older software as Stata and SAS, where the concept of the data set as the basic analytical unit hampers manipulation with model matrices and estimates.

In this paper I reiterated the basic fact for the parametrization of the age–period–cohort model: arbitrary decisions on the allocation of two absolute levels and one drift must be taken in order to report the estimated effects. Arbitrary in this context means unrelated to the model, impossible to derive from data or design. Choice of parametrization should of course not be unrelated to the subject matter.

Since the substance is description of disease rates in populations over time, the major variable is age. Therefore, the reporting of the models should be based on age-specific rates. It is impossible to give universal guidelines as to whether they should be reported as cohort-rates (longitudinal) in

DOI: 10.1002/sim

which case the drift should be included with the cohort-effect, or as period-rates (cross-sectional) in which case the drift should be included with the period effect.

10.1. Recommendations

In summary, the following steps should be taken when describing rates based on observations from a Lexis diagram:

1. Tabulate cases and person-years as detailed as possible, preferably by age, period *and* cohort.
2. Compute risk time from population data using the formulae for risk time in triangles.
3. Use the mean age, period and cohort in each cell of the table as continuous covariates. $a = p - c$ must be met for all analysis units.
4. Use parametric functions to describe the effects. Choose the parametrization (allocation of knots, etc.) carefully, so that relevant features can be captured, but modelling of random noise is avoided.
5. Report estimates of three effects that can be combined to the predicted rates.
6. Age should be the primary variable, report age-specific rates, i.e. include the absolute level with the age-parameters.
7. Make an informed choice of the other aspects of parametrization, and state it clearly. This will include:
   (a) How is the drift extracted.
   (b) Where is the drift allocated.
   (c) How are the RRs for period and cohort fixed.
   (d) What is the interpretation of the age-specific rates.
8. Report estimates as line-graphs with confidence limits.
9. Be careful with firm interpretation of formal tests for period and cohort effects—significant effects may represent clinically irrelevant effects.
10. Do not report goodness of fit tests—they are largely meaningless.

The possible options for parametrization of the model are summarized in Table II.

Table II. Parametrizations for the age–period–cohort model.

Drift extraction:	1: Orthogonal projection	(a) All units the same weight (b) Weights $\propto$ observed information (D) (c) Weights $\propto$ original data (Y)
	2: Equation of two points on period and cohort scales	
Interpretation of effects: Age	Period	Cohort
Longitudinal rates for cohort c_0	Residual RR	RR relative to cohort c_0
Cross-sectional rates for period p_0	RR relative to period p_0	Residual RR
Longitudinal rates for cohort c_0. Fitted rates at $a_0 = p_0 - c_0$	Residual RR relative to p_0	RR relative to cohort c_0
Cross-sectional rates for period p_0. Fitted rates at $a_0 = p_0 - c_0$	RR relative to period p_0	Residual RR relative to c_0

Note: Two steps are needed: fixing the drift parameter, and fixing the reference values.

DOI: 10.1002/sim

The options for model fitting, parametrization and graphical reporting mentioned above are all implemented in the R-package `Epi`, in functions `apc.fit`, `apc.frame` and `apc.lines`. Projections of model matrices is implemented in the function `detrend`. The package is available through CRAN (The Comprehensive R Archive Network, `http://cran.r-project.org`) or at the package home page `http://www.biostat.ku.dk/~bxc/Epi`.

APPENDIX A: PERSON-YEARS IN LEXIS TRIANGLES

The following is based on material from lecture notes by Sverdrup [4]. To my knowledge this has not been published elsewhere, despite its obvious relevance in descriptive epidemiology. The paper by Hoem [21] and the correction note [22] has a reference to a similar result from an earlier version of Sverdrup's notes.

A.1. Census data

First, consider for the sake of simplicity the division of the Lexis-diagram in 1-year classes by age, calendar time and date of birth, and suppose that population figures are available in 1-year classes each year, as will be the situation for most areas where regular censuses are done. The situation is illustrated in Figure 3. The target is to construct estimates of population risk time for each of the areas A and B.

In the following we let a refer to age, p to calendar time (period), and c to date of birth (cohort), and we let $\mathsf{L}_{a,p}$ represent the population size in age a at the beginning of the year p.

If no deaths or migrations occurred in the population, we would have that $\mathsf{L}_{a,p} = \mathsf{L}_{a+1,p+1}$.

In presence of mortality[§] we can at least infer that the survivors $\mathsf{L}_{a+1,p+1}$ have been at risk throughout the year p. Assuming that the persons are uniformly distributed within the age-classes, the average risk time contribution of a survivor will be $\frac{1}{2}$ year to each of the triangles A and B.

In order to work out the contribution of risk time of those dying during the year p, we assume that the deaths are uniformly distributed over A and B.[¶] This means that the total amount of risk time contributed to A and B by those dying in A and B is $(\mathsf{L}_{a,p} - \mathsf{L}_{a+1,p+1}) \times \frac{1}{2}\mathrm{y}$.

Those who die in A contribute no risk time to B. In A their average contribution can be computed by integration over the triangle A. The mean contribution must be calculated as an average w.r.t. to the uniform measure on A. The area of A is $\frac{1}{2}(= \int_{p=0}^{p=1} \int_{a=p}^{a=1} 1 \,\mathrm{d}a \,\mathrm{d}p)$, so the density of the uniform measure is 2.

For simplicity of notation it is assumed that age and date range from 0 to 1 in all of the calculations below. A person dying in age a at date p in A contributes p risk time, so the average is found by integration of the function $f(a, p) = p$ with respect to the uniform measure with density 2 over A, letting a vary from p to 1 and p from 0 to 1

$$\int_{p=0}^{p=1} \int_{a=p}^{a=1} 2p \,\mathrm{d}a \,\mathrm{d}p = \int_{p=0}^{p=1} 2p(1-p) \,\mathrm{d}p = \left[p^2 - \frac{2}{3} p^3 \right]_{p=0}^{p=1} = \frac{1}{3}$$

[§] Immigration and emigration can be treated as negative and positive mortality respectively, and does not alter the results derived here, provided the assumptions made for the mortality pattern also holds for the migration patterns in the population.

[¶] Note that this may a unrealistic assumption for age-classes of length 5 years or more.

DOI: 10.1002/sim

Those who die in B contribute risk time in both A and B. If death occurs in age a at date p the person has contributed $p-a$ person-years in A and a person-years in B.

Hence, the average amount contributed in A is

$$\int_{p=0}^{p=1}\int_{a=0}^{a=p} 2(p-a)\,\mathrm{d}a\,\mathrm{d}p = \int_{p=0}^{p=1} [2pa-a^2]_{a=0}^{a=p}\,\mathrm{d}p = \int_{p=0}^{p=1} p^2\,\mathrm{d}p = \frac{1}{3}$$

and in B

$$\int_{p=0}^{p=1}\int_{a=0}^{a=p} 2a\,\mathrm{d}a\,\mathrm{d}p = \int_{p=0}^{p=1} p^2\,\mathrm{d}p = \frac{1}{3}$$

Under the assumption that the deaths in $\mathsf{A}\cup\mathsf{B}$, $(\mathsf{L}_{a,p}-\mathsf{L}_{a+1,p+1})$ are uniformly distributed, we therefore have the following risk time in A and B

	A	B
Survivors	$\mathsf{L}_{a+1,p+1}\times\frac{1}{2}\mathrm{y}$	$\mathsf{L}_{a+1,p+1}\times\frac{1}{2}\mathrm{y}$
Dead in A	$\frac{1}{2}(\mathsf{L}_{a,p}-\mathsf{L}_{a+1,p+1})\times\frac{1}{3}\mathrm{y}$	
Dead in B	$\frac{1}{2}(\mathsf{L}_{a,p}-\mathsf{L}_{a+1,p+1})\times\frac{1}{3}\mathrm{y}$	$\frac{1}{2}(\mathsf{L}_{a,p}-\mathsf{L}_{a+1,p+1})\times\frac{1}{3}\mathrm{y}$
$\sum$	$(\frac{1}{3}\mathsf{L}_{a,p}+\frac{1}{6}\mathsf{L}_{a+1,p+1})\times 1\mathrm{y}$	$(\frac{1}{6}\mathsf{L}_{a,p}+\frac{1}{3}\mathsf{L}_{a+1,p+1})\times 1\mathrm{y}$

The risk among 0-year olds in year p, born in year p can be computed by requiring that the total risk time among 0-year olds in year p should equal 1y $\times$ the average of the population sizes in age 0 at the beginning and end of year p, i.e. we should use

$$\tfrac{1}{2}(\mathsf{L}_{0,p}+\mathsf{L}_{0,p+1})\times 1\mathrm{y} - (\tfrac{1}{3}\mathsf{L}_{0,p}+\tfrac{1}{6}\mathsf{L}_{1,p+1})\times 1\mathrm{y} = (\tfrac{1}{6}\mathsf{L}_{0,p}+\tfrac{1}{2}\mathsf{L}_{0,p+1}-\tfrac{1}{6}\mathsf{L}_{1,p+1})\times 1\mathrm{y}$$

Another possible estimate is $\frac{1}{2}\mathsf{L}_{0,p+1}$ for those born in year p, disregarding those dead in the year born. Yet another alternative is to take a weighted average of $\frac{1}{2}\mathsf{L}_{0,p+1}$ and half the number of births in the year, $\frac{1}{2}b_p$. Since the mortality is largest in early months, b_p should be given the smallest weight, but the actual weights to use is matter of taste.

A similar procedure can be applied in the last non-open age-class (usually 89). It has little meaning to try to subdivide open age-classes by date of birth.

APPENDIX B: PRACTICALITIES OF PROJECTION

In order to get an estimate of the extracted drift with confidence intervals we need a function that takes P columns of the design matrix produce and a set of $P-2$ columns orthogonal to the constant and the drift w.r.t. some defined inner product.

So let $\mathbf{M}$ be a design matrix, and define the relevant inner product between two columns as

$$\langle \mathbf{m}_j | \mathbf{m}_k \rangle = \sum_i m_{ij} w_i m_{ik}$$

DOI: 10.1002/sim

where we would use either $w_i = 1$, $w_i = D_i$ or $w_i = Y_i$, the total number of cases or person-years observed in unit i of the data set. The task is now to produce a projection of the columns of $\mathbf{M}$ on the orthogonal complement to the two column matrix of the constant and the drift, $[1|p]$ (or $[1|c]$ for the cohort effect).

B.1. Projections in matrix formulation

The projection of a vector $\mathbf{v}$ on the column space of the matrix $\mathbf{X}$ with respect to the usual inner product, is $\mathbf{Pv}$ where

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$$

and the projection on the orthogonal complement is $(\mathbf{I} - \mathbf{P})\mathbf{v}$.

For a general inner product

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum_i x_i w_i y_i = \mathbf{x}^\mathsf{T}\mathbf{W}\mathbf{y}$$

with $\mathbf{W} = \mathrm{diag}(w_i)$, the projection matrix on the column space of $\mathbf{X}$ w.r.t. this inner product is

$$\mathbf{P}_\mathbf{W} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W} \tag{B1}$$

and the projection on the orthogonal complement is $(\mathbf{I} - \mathbf{P}_\mathbf{W})\mathbf{v}$.

B.2. Implementation in R

In the parametrization of the linear trend from the cohort and period effects we use R functions (`ns`, `bs`, ...) to generate model matrices using e.g. natural splines. Let this be $\mathbf{M}$.

Then we project the columns of $\mathbf{M}$ on the orthogonal complement of $[1|p]$ w.r.t. a weighted inner product. This is done using the function `proj.ip`, which is just a translation of the formula (B1) to R (`X` is here playing the role of $[1|p]$)

```
proj.ip <-
function( X, M, orth = FALSE, weight=rep(1,nrow(X)) )
{
 Pp <- solve( crossprod( X * sqrt(weight) ), t( X * weight ) ) %*% M
 PM <- X %*% Pp
 if (orth) PM <- M - PM
 else PM
}
```

When using tabulation of data in very small intervals the resulting data sets can be quite large; for example the example data set with testis cancer in ages 15–65 for the period 1943–1997 has 50 age classes and 54 periods, i.e. $50 \times 54 \times 2 = 5400$ observations for triangles in the Lexis diagram. Therefore, the multiplication $\mathbf{XW}$ is done by multiplying with the *vector* $\mathbf{w}$ (`weight`), using the R-feature of recycling and the column-major storage of matrices. If it had been coded `X %*% diag(weight)`, it would require a square matrix of dimension n (the number of units), which is very large ($5400^2 = 2\,916\,000$ entries). By the same token, the projection matrix $\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}$ is not computed as it would also have this huge dimension. Instead the last

DOI: 10.1002/sim

part of the matrix with projected columns, $(\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{W}\mathbf{M}$ of dimension $p \times p$ is computed first and then multiplied by $\mathbf{X}$ afterwards to produce the $n \times p$ matrix of projected columns. This caution must also exercised if the formulae are to be implemented in other programmes and used in large data sets.

The resulting matrix has the same number of columns as $\mathbf{M}$, so in order to avoid problems with parametrizations, it is shaved down to full rank using `Thin.col`:

```
Thin.col <-
function (X, tol = 1e-06)
{
 QR <- qr(X, tol = tol, LAPACK = FALSE)
 X[, QR$pivot[seq(length = QR$rank)], drop = FALSE]
}
```

These two functions are then used in combination to construct the function `detrend` that de-trends the design matrix $\mathbf{M}$

```
detrend <-
function( M, t, weight=rep(1,nrow(M)) )
{
 Thin.col( proj.ip( cbind( 1, t ), M, orth = TRUE, weight = weight ) )
}
```

Thus, we can find the projection of the design matrix onto the orthogonal of the constant and the period drift, $[1|p]$, by simply `detrend( M, p )`, and it will give the right parametrization for any kind of choice of $\mathbf{M}$.

B.3. Fixing the reference

To get the cohort effect fixed at c_0, a row of the design matrix for cohort corresponding to c_0 is generated. This value need not be a value actually present in the data. When projecting the cohort part of the design matrix on the orthogonal complement of $[1|c]$ the projected (i.e. de-trended value) for the c_0-row can be obtained by amending the design matrix by this row and projecting on the orthogonal complement of $[1|c]$ with respect to an inner product which has weight 0 in the first position. The resulting row is then duplicated to form a matrix of the same size as the design matrix and subtracted from the design matrix. Inside `apc.fit`, the relevant piece of code is

```
xC <- detrend( rbind( Rc, MC ), c(c0,C), weight=c(0,wt) )
MCr <- xC[-1,] - xC[rep(1,nrow(MC)),]
```

where `Rc` is the row of `MC` corresponding to c_0, `MC` is the cohort-design matrix and `C` the vector of cohort midpoints.

B.4. Putting it together

In order to obtain parameters corresponding to log rates by age, a design matrix representing age, including the intercept must be in the model. These will represent log age-specific rates for cohort

DOI: 10.1002/sim

c_0 if the variable $c - c_0$ is put in the model. If the column $c - c_0$ is merged with the de-trended and c_0-centred cohort effect design matrix, this will represent the log RR relative to cohort c_0. Finally, the de-trended period matrix will represent the residual log RR by period.

The estimated age-curve is found by taking the unique rows of the age-part of the design matrix where each row corresponds to an observed age in data. This is then multiplied with the vector of age-parameters to give the curve of estimated log rates. Pre- and post-multiplication on the variance covariance matrix of the age-parameters gives the variances needed to construct confidence limits for the log rates. Finally the log rates are transformed to the rate scale.

The same procedure is used to obtain the RR curves for period and cohort.

REFERENCES

1. Clayton D, Schifflers E. Models for temporal variation in cancer rates. I: Age–period and age–cohort models. *Statistics in Medicine* 1987; **6**:449–467.
2. Clayton D, Schifflers E. Models for temporal variation in cancer rates. II: Age–period–cohort models. *Statistics in Medicine* 1987; **6**:469–481.
3. Keiding N. Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London, Series A* 1990; **332**:487–509.
4. Sverdrup E. *Statistiske metoder ved dødelikhetsundersøkelser*. Statistical Memoirs. Institute of Mathematics, University of Oslo, 1967 (in Norwegian).
5. Tango T. Re: Statistical modelling of lung cancer laryngeal cancer incidence in Scotland 1960–1979. *American Journal of Epidemiology* 1988; **127**(3):677–678.
6. Carstensen B, Keiding N. *Age–Period–Cohort Models*: *Statistical Inference in the Lexis Diagram*. Lecture Notes. Department of Biostatistics, University of Copenhagen, 2004 (http://www.biostat.ku.dk/~bxc/APC/notes.pdf).
7. Osmond C, Gardner MJ. Age, period, and cohort models. Non-overlapping cohorts don't resolve the identification problem. *American Journal of Epidemiology* 1989; **129**(1):31–35.
8. Heuer C. Modelling of time trends and interactions in vital rates using restricted regression splines. *Biometrics* 1997; **53**(1):161–177.
9. Holford TR. Approaches to fitting age–period–cohort models with unequal intervals. *Statistics in Medicine* 2006; **25**:977–993.
10. Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics* 1983; **39**:311–324.
11. Richiardi L, Bellocco R, H-Adami O, Torrång A, Barlow L, Hakulinen T, Rahu M, Stengrevics A, Storm H, Tretli S, Kurtinaitis J, Tyczynski JE, Akre O. Testicular cancer incidence in eight northern European countries: secular and recent trends. *Cancer Epidemiology Biomarkers and Prevention* 2004; **13**(12):2157–2166.
12. Ajdacic-Gross V, Bopp M, Gostynski M, Lauber G, Gutzwiller F, Rossler W. Age–period–cohort analysis of Swiss suicide data 1881–2000. *European Archives of Psychiatry and Clinical Neuroscience* 2005; **256**(4): 207–214.
13. Bruno G, Merletti F, Biggeri A, Cerutti F, Grosso N, De Salvia A, Vitali E, Pagano G. Increasing trend of type I diabetes in children young adults in the province of Turin (Italy). Analysis of age, period and birth cohort effects from 1984 to 1996. *Diabetologia* 2001; **44**(1):22–25.
14. Feltbower RG, McKinney PA, Parslow RC, Stephenson CR, Bodansky HJ. Type 1 diabetes in Yorkshire, U.K.: Time trends in 0–14 and 15–29-year-olds, age at onset and age–period–cohort modelling. *Diabetic Medicine* 2003; **20**(6):437–441.
15. Pundziute-Lyckå A, Dahlquist G, Nyström L, Arnqvist H, Björk E, Blohmé E, Bolinder J, Eriksson JW, Sundkvist G, Östman J, Swedish Childhood Diabetes Group Study. The incidence of type I diabetes has not increased but shifted to a younger age at diagnosis in the 0–34 years group in Sweden 1983 to 1998. *Diabetologia* 2002; **45**:773–791.
16. Rewers M, Stone RA, La Porte RE, Drash AL, Becker DJ, Walczak M, Kuller LH. Poisson regression modelling of temporal variation in incidence of childhood insulin-dependent diabetes mellitus in Allegheny county, Pennsylvania and Wielkopolska, Poland, 1970–1985. *American Journal of Epidemiology* 1989; **129**(3):569–581.
17. Boyle P, Robertson C. Statistical modelling of lung cancer and laryngeal cancer incidence in Scotland 1960–1979. *American Journal of Epidemiology* 1987; **125**(4):731–744.

DOI: 10.1002/sim

18. Robertson C, Boyle P. Age, period and cohort models: the use of individual records. *Statistics in Medicine* 1986; **5**:527–538.
19. Robertson C, Boyle P. Age–period–cohort analysis of chronic disease rates. I: Modelling approach. *Statistics in Medicine* 1998; **17**:1305–1323.
20. Tyczynski JE, Hill TD, Berkel HJ. Why do postmenopausal African-American women not benefit from overall breast cancer mortality decline? *Annals of Epidemiology* 2005; **16**(3):180–190.
21. Hoem JM. Fertility rates and reproduction rates in a probabilistic setting. *Biométrie-Praximétrie* 1969; **10**: 38–66.
22. Hoem JM. Correction note. *Biométrie-Praximétrie* 1970; **11**:20.

DOI: 10.1002/sim