STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:1557–1564

LETTER TO THE EDITOR

Comments on 'Age–period–cohort models for the Lexis diagram'

by Carstensen B. *Statistics in Medicine* 2007; **26**:3018–3045

From: *Joachim Rosenbauer and Klaus Strassburger*
Institute of Biometrics and Epidemiology,
German Diabetes Centre, *Leibniz Institute for Diabetes Research*
at Heinrich-Heine-University Düsseldorf,
Auf'm Hennekamp 65,
D-40225 Düsseldorf, *Germany*

In a recent article, Carstensen [1] gives a concise and excellent guide to the analysis of disease rates from a Lexis diagram by the age–period–cohort model. Starting from the classical approach, modelling age, period and cohort effects by categorical variables (class variables, factors), the author particularly considers the 'natural approach' to model effects in continuous time by parametric smooth functions (e.g. B-splines) of means of triangular subsets of the Lexis diagram. This approach saves fitting two separate models to data classified by age, period and cohort, unlike in the classical factor model [2]. Practical recommendations for parameterization of models and presentation of estimated effects are given and an implementation of the methods for **R** is introduced.

It is an essential precondition for the application of the proposed methods that disease cases and person time at risk (person-years) can be tabulated with respect to age, calendar time (period) and date of birth (cohort) for a triangular subset of the Lexis diagram (Figure 1). While cases from registers (dates of birth and diagnosis are typically known) can easily be allocated to $1 \times 1 \times 1$-year triangles of the Lexis diagram, the availability of population figures will normally be the limiting factor.

In his article, Carstensen [1] presents formulas for the estimation of population risk time for triangular age–period–cohort subsets of the Lexis diagram based on population data in 1-year age classes for each calendar year, which are available for most countries. The derivation of the formulas traces back to lecture notes by Sverdrup [3], an earlier version of which has also been referenced by Hoem [4, 5].

However, the formulas for population risk time presented are obviously flawed. Thus, results based on the proposed modelling approach are potentially inaccurate, although there may be only slight bias in the estimation of person-years—bias will be largest in older age groups due to higher mortality.

In this letter, we derive corrected formulas for the estimation of population risk time in triangular subsets of the Lexis diagram.

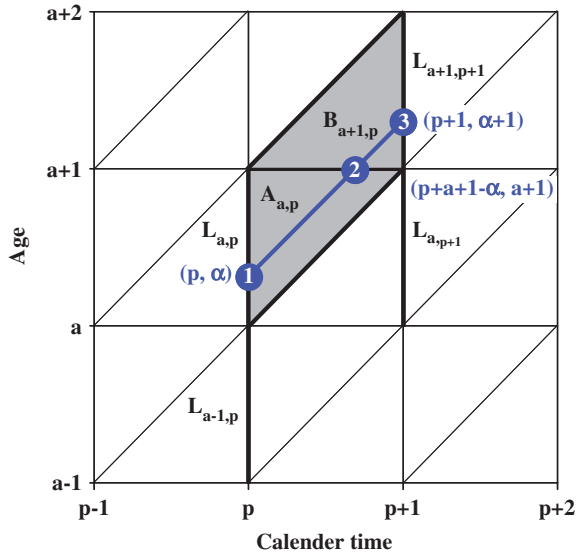

Figure 1. Lexis diagram: the vertical thick lines show the population figures at the beginning of year p in 1-year age classes $a-1$ ($L_{a-1,p}$) and a ($L_{a,p}$), and at the beginning of year $p+1$ in 1-year age classes a ($L_{a,p+1}$) and $a+1$ ($L_{a+1,p+1}$) years. Based on these figures, population risk time can be estimated for the triangular subsets $A_{a,p}$ and $B_{a+1,p}$. The line between points 1 and 2 corresponds to a potential life line section of a person in age α at the beginning of year p. α determines the maximum potential risk time contributed to subsets $A_{a,p}$ and $B_{a+1,p}$. Time elapsed between crossing points of the life line with date p (point 1) and age $a+1$ (point 2) is $(a+1-\alpha)$, and that between crossing points with age $a+1$ (point 2) and date $p+1$ (point 3) is $(\alpha-a)$.

ESTIMATION OF PERSON-YEARS IN TRIANGULAR SUBSETS OF THE LEXIS-DIAGRAM

In short, the Lexis diagram is an age by calendar time coordinate system in which individual lives are represented by line segments with unit slope. It is supposed that the Lexis diagram is subdivided into 1-year classes with respect to age, calendar time and date of birth, and that population data are available stratified by age and calendar year in 1-year classes. The situation is illustrated in Figure 1 (according to Carstensen). The aim is to estimate population risk time for the triangular subsets $A_{a,p}$ and $B_{a+1,p}$ of the Lexis diagram.

Following Carstensen's notation, we let a refer to 1-year age classes and p to calendar years (period). The population size in age class a at the beginning of the year p is denoted by $L_{a,p}$.

When assuming ages of the persons $L_{a,p}$ to be equally distributed within age class a and deaths $(L_{a,p}-L_{a+1,p+1})$ to be equally distributed over year p* (corresponding to subset $A_{a,p} \cup B_{a+1,p}$)—thus half of these deaths occur in $A_{a,p}$ and half in $B_{a+1,p}$—risk time contributions of the persons

*It has to be emphasized that formulas for deceased persons in Table I presume that deaths, i.e. person risk times are equally distributed in triangular subsets (see also [1, p. 3040 and Appendix A]), and not that mortality rates are constant in triangular subsets, as is misleadingly stated on p. 3023 in [1].

Table I. Estimates of population risk time for triangular subsets of the Lexis diagram.

	Risk time in $A_{a,p}$	Risk time in $B_{a+1,p}$
Survivors ($L_{a+1,p+1}$)	$L_{a+1,p+1} \cdot \frac{1}{2} py$	$L_{a+1,p+1} \cdot \frac{1}{2} py$
Deceased in $A_{a,p}$	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1}) \cdot \frac{1}{4} py$	—
Deceased in $B_{a+1,p}$	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1}) \cdot \frac{1}{2} py$	$\frac{1}{2}(L_{a,p} - L_{a+1,p+1}) \cdot \frac{1}{4} py$
All persons ($L_{a,p}$)	$(\frac{3}{8} L_{a,p} + \frac{1}{8} L_{a+1,p+1}) \cdot 1 py$	$(\frac{1}{8} L_{a,p} + \frac{3}{8} L_{a+1,p+1}) \cdot 1 py$

py, Person-year.

$L_{a,p}$ to $A_{a,p}$ and $B_{a+1,p}$ are estimated stratified by survival status with respect to year p according to the formulas in Table I. A detailed derivation is given in the Appendix.

According to the estimates in Table I, the total risk time in age class a and year p—corresponding to subset $A_{a,p} \cup B_{a,p}$—is best estimated by $\frac{1}{8} L_{a-1,p} + \frac{3}{8} L_{a,p} + \frac{3}{8} L_{a,p+1} + \frac{1}{8} L_{a+1,p+1}$ person-years and not by $\frac{1}{2} L_{a,p} + \frac{1}{2} L_{a,p+1}$ person-years, as commonly done.

The risk time of 0-year olds in year p born in year p can be approximated by subtracting the risk time in $A_{0,p}$ from the available estimate of the risk time among 0-year olds in year p (subset $A_{0,p} \cup B_{0,p}$), i.e.

$$\begin{aligned} &\tfrac{1}{2}(L_{0,p} + L_{0,p+1}) - (\tfrac{3}{8} L_{0,p} + \tfrac{1}{8} L_{1,p+1}) \text{ person-years} \\ &\quad = \tfrac{1}{8} L_{0,p} + \tfrac{1}{2} L_{0,p+1} - \tfrac{1}{8} L_{1,p+1} \text{ person-years} \end{aligned}$$

Note that, according to Carstensen's formulas, the average risk time in $A_{a,p}$ was estimated as $\frac{1}{3}$ person-year for both persons dying in $A_{a,p}$ and persons dying in $B_{a+1,p}$. This obviously cannot be valid, because persons dying in $B_{a+1,p}$ have been under risk throughout subset $A_{a,p}$ but not throughout $B_{a+1,p}$, and therefore contributed risk time to $A_{a,p}$ must exceed the risk time in $B_{a+1,p}$, like in our corrected formulas. The flaw in Carstensen's derivation is that he presumes a uniform two-dimensional measure over triangular subsets $A_{a,p}$ and $B_{a+1,p}$; thus he disregards the fact that maximum potential risk time of a person in $A_{a,p}$ and $B_{a+1,p}$ is determined by the person's age at the beginning of year p. In fact, risk times have to be properly considered conditionally on the age at entrance into $A_{a,p}$, according to 'life lines' in the Lexis diagram.

APPENDIX

Derivation of risk time

Let T_S denote the risk time of a person in a subset $S \in \{A_{a,p}, B_{a+1,p}, A_{a,p} \cup B_{a+1,p}\}$ (Figure 1). Then, the *average risk time* of persons $L_{a,p}$ can formally be estimated as the expectation of the risk time in subset S (denoted by $E(T_S)$). $E(T_S)$ itself can be calculated as an expectation of the

conditional expectation of the risk time with respect to the age α, $a \leqslant \alpha \leqslant a+1$, at the beginning of year p (denoted by $E(T_S|\alpha)$), that is by integration of $E(T_S|\alpha)$ with respect to the uniform measure over the 1-year age class 'a', when assuming age α of persons $L_{a,p}$ at the beginning of year p to be equally distributed within the 1-year age class 'a' (pdf: $f(\alpha)=1$, $a \leqslant \alpha \leqslant a+1$):

$$E(T_S) = E(E(T_S|\alpha, a \leqslant \alpha \leqslant a+1)) = \int_a^{a+1} E(T_S|\alpha) \cdot f(\alpha)\, \mathrm{d}\alpha = \int_a^{a+1} E(T_S|\alpha)\, \mathrm{d}\alpha$$

The conditional risk time $E(T_S|\alpha)$ that a person contributes to $S \in \{A_{a,p}, B_{a+1,p}, A_{a,p} \cup B_{a+1,p}\}$ and thus the expectation of $E(T_S|\alpha)$ in the subset depends on whether the person survives year p or dies in subsets $A_{a,p}$ or $B_{a+1,p}$. In each case, the conditional risk time T_S of a person is assumed to be equally distributed over the respective subset of the Lexis diagram (pdf of T_S denoted as $g(t|\alpha)$, $0 \leqslant t \leqslant 1$, $a \leqslant \alpha \leqslant a+1$).

Risk time of survivors of year p

Clearly, the survivors $L_{a+1,p+1}$ of year p have been at risk throughout the year p. Thus, under the above assumptions, the average risk time contribution of survivors to each of the triangular subsets $A_{a,p}$ and $B_{a+1,p}$ will be $\frac{1}{2}$ year. The total risk time for both subsets $A_{a,p}$ and $B_{a+1,p}$ will be $L_{a+1,p+1} \cdot \frac{1}{2}$ person-years.

Formally, the risk time contribution of a survivor of year p, who has been in age α, $a \leqslant \alpha \leqslant a+1$, at the beginning of year p, to subset $A_{a,p}$ is $(a+1-\alpha)$. Therefore, the *average risk time in* $A_{a,p}$ is

$$\int_a^{a+1} E(T_{A_{a,p}}|\alpha)\, \mathrm{d}\alpha = \int_a^{a+1} (a+1-\alpha)\, \mathrm{d}\alpha = \int_0^1 (1-\alpha)\, \mathrm{d}\alpha = \frac{1}{2} \text{ person-year}$$

and the risk time contribution to $B_{a+1,p}$ is $(\alpha - a)$, giving the average risk time in $B_{a+1,p}$ as

$$\int_a^{a+1} E(T_{B_{a+1,p}}|\alpha)\, \mathrm{d}\alpha = \int_a^{a+1} (\alpha - a)\, \mathrm{d}\alpha = \int_0^1 \alpha\, \mathrm{d}\alpha = \frac{1}{2} \text{ person-year}$$

Risk time of deceased in year p

Deceased in $A_{a,p} \cup B_{a+1,p}$. Since risk time $T_{A_{a,p} \cup B_{a+1,p}}$ is assumed to be equally distributed in $A_{a,p} \cup B_{a+1,p}$, the respective pdf is given by $g(t|\alpha)=1$, $0 \leqslant t \leqslant 1$, $a \leqslant \alpha \leqslant a+1$. Thus, the *average risk time* of the deceased of persons $L_{a,p}$ *in* $A_{a,p} \cup B_{a+1,p}$ is

$$\int_a^{a+1} E(T_{A_{a,p} \cup B_{a+1,p}}|\alpha)\, \mathrm{d}\alpha = \int_a^{a+1} \left(\int_0^1 t \cdot g(t|\alpha)\, \mathrm{d}t \right) \mathrm{d}\alpha = \int_a^{a+1} \left(\int_0^1 t \cdot 1\, \mathrm{d}t \right) \mathrm{d}\alpha$$

$$= \int_0^1 \left(\int_0^1 t\, \mathrm{d}t \right) \mathrm{d}\alpha = \int_0^1 \frac{1}{2}\, \mathrm{d}\alpha = \frac{1}{2} \text{ person-year}$$

Deceased in $A_{a,p}$. Persons dying in $A_{a,p}$ contribute no risk time to $B_{a+1,p}$.

For a person who entered subset $A_{a,p}$ at age α, $a \leqslant \alpha \leqslant a+1$, and died in $A_{a,p}$ the risk time is assumed to be equally distributed on the interval $[0, a+1-\alpha]$, giving the respective pdf as

$g(t|\alpha) = 1/(a+1-\alpha)$, $0 \leqslant t \leqslant 1$, $a \leqslant \alpha \leqslant a+1$. Hence, the *average risk time in* $A_{a,p}$ is given by

$$\int_a^{a+1} E(T_{A_{a,p}}|\alpha)\,\mathrm{d}\alpha = \int_a^{a+1} \left(\int_0^{a+1-\alpha} t \cdot g(t|\alpha)\,\mathrm{d}t \right) \mathrm{d}\alpha = \int_a^{a+1} \left(\int_0^{a+1-\alpha} t \cdot \frac{1}{(a+1-\alpha)}\,\mathrm{d}t \right) \mathrm{d}\alpha$$

$$= \int_0^1 \left(\int_0^{1-\alpha} t \cdot \frac{1}{(1-\alpha)}\,\mathrm{d}t \right) \mathrm{d}\alpha = \int_0^1 \frac{1-\alpha}{2}\,\mathrm{d}\alpha = \frac{1}{4} \text{ person-year}$$

Deceased in $B_{a+1,p}$. A person who entered $A_{a,p}$ at age α, $a \leqslant \alpha \leqslant a+1$, and died in $B_{a+1,p}$ has been under risk throughout $A_{a,p}$ and contributes a risk time of $(a+1-\alpha)$ person-year in $A_{a,p}$. Thus, the *average risk time in* $A_{a,p}$ is given by

$$\int_a^{a+1} E(T_{A_{a,p}}|\alpha)\,\mathrm{d}\alpha = \int_a^{a+1} (a+1-\alpha)\,\mathrm{d}\alpha = \int_0^1 (1-\alpha)\,\mathrm{d}\alpha = \frac{1}{2} \text{ person-year}$$

The risk time of such a person in $B_{a+1,p}$ varies on $[0, \alpha - a]$ and is assumed to be equally distributed, giving the respective pdf as $g(t|\alpha) = 1/(\alpha - a)$, $0 \leqslant t \leqslant 1$, $a \leqslant \alpha \leqslant a+1$. Hence, the *average risk time in* $B_{a+1,p}$ is estimated by

$$\int_a^{a+1} E(T_{A_{a,p}}|\alpha)\,\mathrm{d}\alpha = \int_a^{a+1} \left(\int_0^{\alpha-a} t \cdot g(t|\alpha)\,\mathrm{d}t \right) \mathrm{d}\alpha = \int_a^{a+1} \left(\int_0^{\alpha-a} t \cdot \frac{1}{\alpha - a}\,\mathrm{d}t \right) \mathrm{d}\alpha$$

$$= \int_0^1 \left(\int_0^{\alpha} t \cdot \frac{1}{\alpha}\,\mathrm{d}t \right) \mathrm{d}\alpha = \int_0^1 \frac{\alpha}{2}\,\mathrm{d}\alpha = \frac{1}{4} \text{ person-year}$$

REFERENCES

1. Carstensen B. Age–period–cohort models for the Lexis diagram. *Statistics in Medicine* 2007; **26**(15):3018–3045. DOI: 10.1002/sim.2764.
2. Osmond C, Gardner MJ. Age, period, cohort models. Non-overlapping cohorts don't resolve the identification problem. *American Journal of Epidemiology* 1989; **129**(1):31–35.
3. Sverdrup E. Statistiske metoder ved dødelikhetsundersøkelser (in Norwegian). *Statistical Memoirs*. Institute of Mathematics, University of Oslo, 1967.
4. Hoem JM. Fertility rates and reproduction rates in a probabilistic setting. *Biométrie-Praximétrie* 1969; **10**(1):38–66.
5. Hoem JM. Correction note. *Biométrie-Praximétrie* 1970; **11**(1):20.

Published online 11 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.3041

AUTHOR'S REPLY

Age–period–cohort models for the Lexis diagram, *Statistics in Medicine* 2007; **26**:3018–3045

I thank Joachim Rosenbauer and Klaus Strassburger (R&S) for their interest in the paper and the derivations of the population risk time [1].

1. ASSUMPTIONS ABOUT DISTRIBUTION OF DEATHS

The core of their argument is that computation of risk time in age × period × cohort subsets of a Lexis diagram should be done conditional to the age at entry into the upper triangle sets, termed A both in my paper and their letter.

R&S derive precisely what I in the paper only asserted by handwaving, namely that the contribution of risk time in the sets A and B from persons in age class a at calendar time p, who survive to time $p+1$ (and hence at that time in age class $a+1$), is on average $\frac{1}{2}$ year in each. Underlying the calculation is the perfectly reasonable assumption that the distribution of ages at time p among those surviving to time $p+1$ is uniform on the interval $[a, a+1]$—the integration is with respect to age at time p using the uniform measure on $[a, a+1]$. This assumption is only approximately the same as the assumption of a uniform age distribution for *all* persons alive at p ($L_{a,p}$) if mortality rates do not vary dramatically by age. The same arguments will hold in the case of computing the risk time for all those who die in the set $\mathsf{A} \cup \mathsf{B}$.

But R&S use exactly the same argument and assumptions when they compute the risk time separately in each of the sets A and B from persons who die in these sets, particularly the assumption of uniform age distribution at time p. However in the case of deaths this is highly untenable. Given that a person has died in the set A makes it much more likely that the person entered at an early age and had a long exposure.

Hence, the computation by R&S is correct but relies on different assumptions, which I consider counterintuitive.

2. A SMALL SIMULATION STUDY

To illustrate this I carried out a small simulation study as follows: R&S use the assumption that the age at time p for persons who die in either A or B is uniformly distributed over the interval $[a, a+1]$, and in order to complete the computations the additional assumption that the time of death given entry age α is uniformly distributed on $[p, p+1-\alpha]$. This is sufficient to simulate a number of deaths in each of A and B. Once this is done, the empirical distribution of ages at p and of the risk time in each of the sets can be computed. Similarly, the assumptions that I use in my paper, uniform distribution of deaths over $\mathsf{A} \cup \mathsf{B}$, are easily simulated and the same computations on the simulated sample carried out.

The difference between the two approaches is illustrated in Figure 1. The top part represents 800 deaths in each of A and B with ages at p uniformly distributed and deaths uniformly distributed within the possible follow-up (R&S assumptions). The lower half represents 1600 deaths uniformly distributed over A and B (my assumptions).

The assumptions that R&S make imply a very odd clumping of deaths in the corners of A and B, which is indeed difficult to find a justification for. The assumptions that I make in the paper induces a highly skewed distributions of age at p, given death in either A or B, which however is a perfectly sensible consequence.

The empirical means of the risk time for the deaths based on the simulated samples are shown too, demonstrating that the formulae derived by R&S and me are actually reproduced in the simulations.

R&S note that in my derivation persons who die in B contribute on average the same amount of risk time in A and B, which they find odd since anyone who dies in B has lived through A.

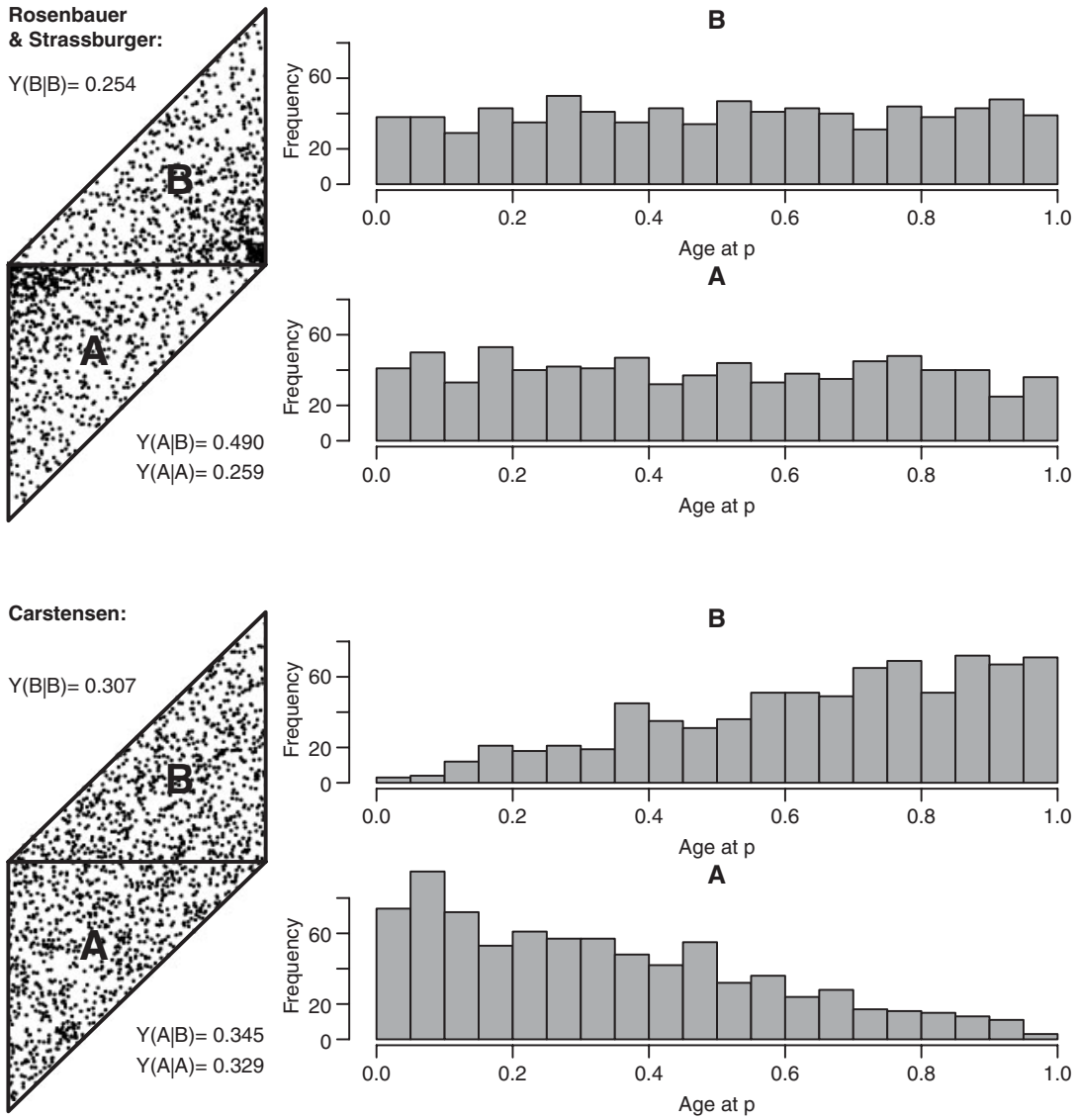

Figure 1. Results from simulation of deaths by two approaches. Top half: uniform distribution of age at start—a very strange assumption and not recommendable. Bottom half: uniform distribution of deaths over A and B—reasonable approximation in practise. The R program that does the simulation and the plot is available as http://staff.pubhealth.ku.dk/~bxc/APC/R/Rosenbauer-Strassburger.R.

But not all persons dying in B have spent the same time in A; the shorter the time spent in A, the longer the time spent in B. That is the intuitive explanation—the mathematical one is in my paper.

3. CONCLUSION

My derivation in the paper [2] is correct and based on demographically sensible assumptions; the derivation by R&S is correct but based on assumptions that are highly unlikely to be relevant in any practical circumstances.

Therefore, the formulae given in my paper [2] are those that should be used in practice.

B. CARSTENSEN
Steno Diabetes Center, *Niels Steensens Vej 2*
DK.2820 Gentofte, *Denmark*
E-mail: bxc@steno.dk

REFERENCES

1. Rosenbauer J, Strassburger K. Letter: Re: Age–period–cohort models for the Lexis diagram. *Statistics in Medicine* 2007; **26**:3018–3045.
2. Carstensen B. Age–period–cohort models for the Lexis diagram. *Statistics in Medicine* 2007; **26**:3018–3045.

Published online 4 September 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.3058

