DISCUSSION PAPER PI-0814

Modelling and Management of Mortality Risk: A Review

Andrew J.G. Cairns, David Blake and Kevin Dowd

October 2008

ISSN 1367-580X

The Pensions Institute
Cass Business School
City University
106 Bunhill Row London
EC1Y 8TZ
UNITED KINGDOM

http://www.pensions-institute.org/

Scandinavian Actuarial Journal, 2008, 2–3, 79–113

Taylor & Francis
Taylor & Francis Group

Original Article

Modelling and management of mortality risk: a review

ANDREW J. G. CAIRNS*, DAVID BLAKE† and KEVIN DOWD‡

*Maxwell Institute for Mathematical Sciences, and Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, UK
†Pensions Institute, Cass Business School, London, UK
‡Centre for Risk & Insurance Studies, Nottingham University Business School, Nottingham, UK

(*Accepted 28 January 2008*)

In the first part of the paper, we consider the wide range of extrapolative stochastic mortality models that have been proposed over the last 15–20 years. A number of models that we consider are framed in discrete time and place emphasis on the statistical aspects of modelling and forecasting. We discuss how these models can be evaluated, compared and contrasted. We also discuss a discrete-time market model that facilitates valuation of mortality-linked contracts with embedded options. We then review several approaches to modelling mortality in continuous time. These models tend to be simpler in nature, but make it possible to examine the potential for dynamic hedging of mortality risk. Finally, we review a range of financial instruments (traded and over-the-counter) that could be used to hedge mortality risk. Some of these, such as mortality swaps, already exist, while others anticipate future developments in the market.

Keywords: Stochastic mortality models; Short-rate models; Market models; Cohort effect; SCOR market model; Mortality-linked securities; Mortality swaps; q-forwards

1. Introduction

The twentieth century has seen significant improvements in mortality rates. Figure 1 demonstrates this for selected ages for English and Welsh males. To facilitate understanding, these rates have been plotted on a logarithmic scale, and with the same relative range of values on the vertical axis (that is, the maximum on each scale is 12 times the minimum). From these plots we can extract the following stylised facts which apply to most developed countries (for example, Western Europe, Scandinavia, North America):

- Mortality rates have fallen dramatically at all ages.
- At specific ages, improvement rates seem to have varied over time with significant improvements in some decades and almost no improvements in other decades. For example, for English and Welsh males, the age 25 rate improved dramatically before 1960 and then levelled off; at age 65 the opposite was true.

Corresponding author. E-mail: a.j.g.cairns@hotmail.co.uk

Scandinavian Actuarial Journal
ISSN 0346-1238 print/ISSN 1651-2030 online
http://www.tandf.no/saj
DOI: 10.1080/03461230802173608

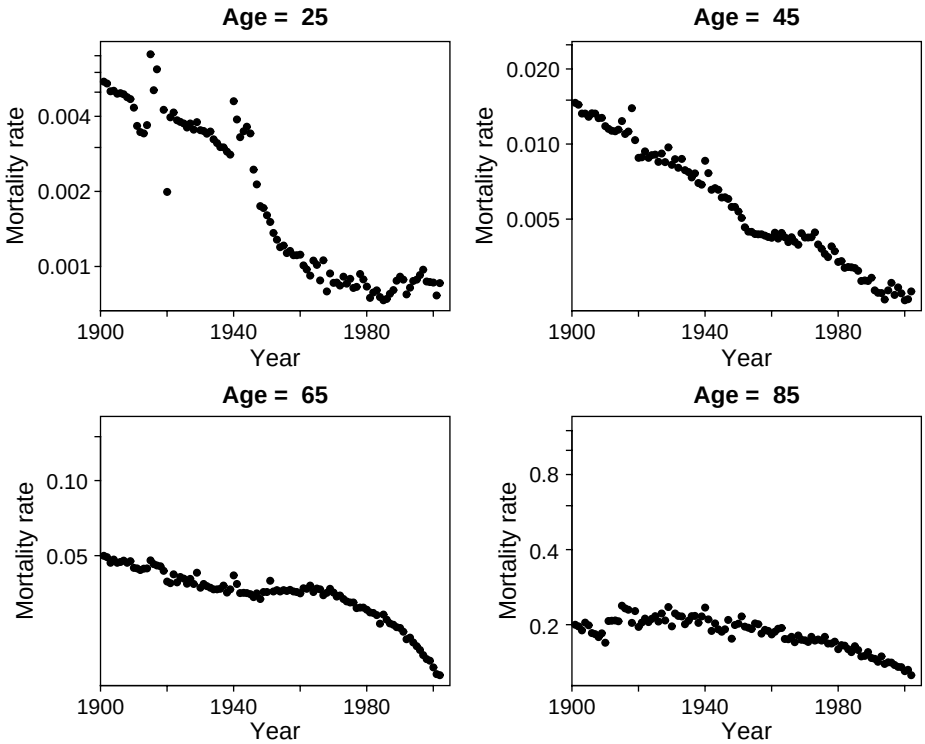

Figure 1. England and Wales males: Mortality rates at ages 25, 45, 65 and 85. Rates are plotted on a logarithmic scale, with the same relative range of values in each plot. The age-25 rate for 1918 is off the scale.

- Improvement rates have been significantly different at different ages. For example, for English and Welsh males, the age 45 improvements have been much higher than the age 85 improvements.
- Aggregate mortality rates reveal significant volatility from one year to the next. This is especially true at both young ages, where the numbers of deaths are relatively small, and high ages. At high ages, the numbers of deaths are, of course, high, so we infer that there are genuine annual fluctuations in the true *underlying* rates of mortality at high ages. These fluctuations might, for example, be due to the incidence of winter flu epidemics or summer heat waves in some years causing increased mortality amongst the old and frail in the population in those years.

From a statistical perspective, we might also add a further point. Since rates of improvement have varied over time and have been different at different ages, there will be considerable uncertainty in forecasting what rates of improvement will be in the future at different ages.

In some, but not all countries, an additional observation has been made that patterns of improvement are linked to year of birth (see, for example, Willets (2004) and Richards

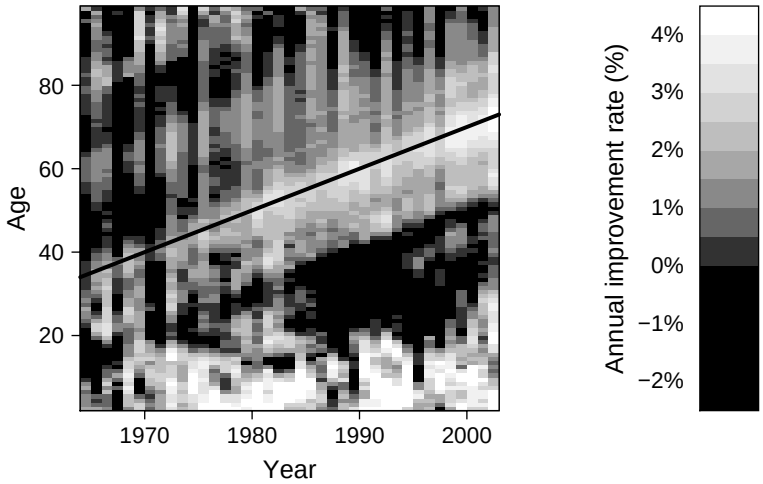

Figure 2. Improvement rates in mortality for England and Wales by calendar year and age relative to mortality rates at the same age in the previous year. Black cells imply that mortality is deteriorating; medium grey small rates of improvement, and white strong rates of improvement. The black diagonal line follows the progress of the 1930 cohort. (Source: Cairns *et al.* (2007))

et al. (2006)). In Figure 2, we have plotted 5-year-average rates of improvement (that is, $-0.2 \log[q(t,x)/q(t-5,x)]$, where $q(t, x)$ is the mortality rate at age x in year t). In this plot, we can see that there are clear groupings of high improvement rates and of low improvement rates. We have specifically highlighted the cohort born around 1930 by a solid black diagonal line. The shading along this diagonal is clearly lighter than other diagonals indicating that individuals born around that time (1925–1935) have experienced rather larger improvements in mortality compared with people born before or after. Possible explanations for this 'golden' cohort include a healthy diet in the 1940s and early 1950s (or, to be more precise, shortages of unhealthy food), and the introduction of the National Health Service. In more general terms, it is easy to detect diagonals above age 30 running through Figure 2, and this is indicative of year of birth potentially being an important factor in projecting future mortality rates for England and Wales males. Also in Figure 2, we can see strong reductions in mortality amongst the very young, and a deterioration in mortality amongst males in their 20s during the late 1980s and 1990s (due to the AIDS epidemic).

1.1. Systematic and unsystematic mortality

It is appropriate at this point to discuss mortality in a more scientific setting. We will define $q(t, x)$ to be the underlying aggregate mortality rate in year t at age x. This is an unobservable rate. What we do observe depends on how, for example, national statistics offices record deaths and population sizes. However, in many countries we observe the crude death rate, $m_c(t, x)$, which is the number of deaths, $D(t, x)$, aged x last birthday at

the date of death, during year t, divided by the exposure, $E(t, x)$ (the average population aged x last birthday during year t).[1]

The uncertainty in future death rates can be divided into two components:

- *Unsystematic mortality risk*. Even if the true mortality rate is known, the number of deaths, $D(t, x)$, will be random. The larger the population, the smaller the unsystematic mortality risk (as a result of the pooling of offsetting risks, i.e., diversification).
- *Systematic mortality risk*. This is the undiversifiable component of mortality risk that affects all individuals in the same way. Specifically, forecasts of mortality rates in future years are uncertain.

1.2. Life assurers and pension plans

The modelling and management of systematic mortality risk are two of the main concerns of large life assurers and pension plans and we will focus in this paper on these two concerns:

- Modelling:
  - What is the best way to forecast future mortality rates and to model the uncertainty surrounding these forecasts?
  - How do we value risky future cashflows that depend on future mortality rates?
- Management:
  - How can this risk be actively managed and reduced as part of an overall strategy of efficient risk management?
  - What hedging instruments are easier to price than others?

In order to address these questions, we require the following: first, we need suitable stochastic mortality models capable of forecasting the distribution of future mortality rates which will help with both quantifying and pricing mortality risk; second, we need suitable vehicles for managing or transferring mortality risk. In addition to focusing on these two requirements, we also review the rapidly developing literature in this field, as well as present some new ideas and models.

In Section 2, we introduce some of the basic notation that is required in the subsequent Sections 3–5 where we discuss specific approaches to modelling. In Section 1, before looking at particular models, we review the range of criteria that can be used to evaluate different models. We then move on, in Sections 4 and 5, to discuss discrete-time and continuous-time models. We develop some new insights as well as some new models (including a new model with a cohort effect, a generalization of the Olivier–Smith market model and further thoughts on the survivor credit offered rate (SCOR) market model). Section 6 reviews existing and potential market instruments that might be used to help manage mortality risk. Section 7 concludes.

[1] In our notation the subscript c in $m_c(t, x)$ distinguishes the crude or actual death rate from the underlying or expected death rate.

2. Basic building blocks

For modelling, we need to define some general notation.

2.1. Mortality rates

$q(t, x)$ is the underlying probability that an individual aged exactly x at time t will die before time $t+1$. The period t to $t+1$ will also be referred to as 'year t'. $q(t, x)$ is normally only defined for integer values of x and t, and is only observable after time $t+1$.

2.2. The instantaneous force of mortality

$\mu(t, x)$ is the underlying force of mortality at time t and age x for real t and x.

We have the relationship

$$q(t,x) = 1 - \exp\left[-\int_t^{t+1} \mu(u, x-t+u)du\right]$$

which we treat as being observable only after time $t+1$.

2.3. The survivor index

The survivor index is defined as

$$S(t,x) = \exp\left[-\int_0^t \mu(u, x+u)du\right].$$

An informal way of thinking about $S(t, x)$ is that it represents the proportion of a large population aged exactly x at time 0 who survive to age $x+t$ at time t. More formally, let $\mathcal{M}_t$ be the filtration generated by the whole of the instantaneous force of mortality curve (that is, covering all ages) up to time t, so that $S(t, x)$ is $\mathcal{M}_t$-measurable. Now consider a single individual aged exactly x at time 0. Let $I(t)$ be the indicator random variable that is equal to 1 if the individual is still alive at time t and 0 if he is dead.

Then $Pr[I(t) = 1|\mathcal{M}_t] = S(t, x)$. Further, $Pr[I(t) = 1|\mathcal{M}_u] = S(t, x)$ for all $u \geq t$.

Remark (See, for example, Biffis *et al.* (2006)) *The filtration $\mathcal{M}_t$ informs us about the underlying mortality dynamics only. It does not tell us about the times of death of individuals in the population. We will denote by $\mathcal{M}_t^*$ the augmented filtration that is generated by both $\mu(t, x)$ and the times of death of the individuals in the population.*

2.4. Spot and forward survival probabilities

Our remark about $u \geq t$ in subsection 2.3 raises the question as to what happens if $u < t$. For $0 < u < t$,

$$Pr[I(t) = 1|\mathcal{M}_u] = E[S(t,x)|\mathcal{M}_u] = S(u,x)E\left[\frac{S(t,x)}{S(u,x)}\middle|\mathcal{M}_u\right].$$

This leads us to the definition of the spot survival probabilities:

$$p(u,t,x) = Pr[I(t) = 1 | I(u) = 1, \mathcal{M}_u]$$

so that $Pr[I(t) = 1|\mathcal{M}_u] = S(u,x)\,p\,(u,t,x)$.

Finally, this leads us to define forward survival probabilities for $T_0 < T_1$:

$$p(t,T_0,T_1,x) = Pr[I(T_1) = 1 | I(T_0) = 1, \mathcal{M}_t]. \tag{1}$$

The name forward survival probability suggests that $t \leq T_0$, but, in fact, the probability is well defined for any t.

Remark *It is important to keep in mind that the spot and forward survival probabilities, $p(t, T, x)$ and $p(t, T_0, T_1, x)$ concern individuals aged x at time 0 and not at time t.*

2.5. Real-world and risk-neutral probabilities

In the remarks above, we have implicitly assumed that probabilities and expectations have been calculated under the true or real-world probability measure, P. When it comes to pricing, we will often use an artificial, risk-neutral probability measure, Q. The extent to which P and Q differ depends, for example, upon how much of a premium life assurers and pension plans would be prepared to pay to hedge their systematic and possibly non-systematic mortality risks, giving rise to the concept of a market price of risk.

Most previous research (see, for example, Cairns *et al.* (2006a), and references therein) assumes that there is a market price of risk for systematic mortality risk only and that non-systematic risk, being diversifiable, goes unrewarded. Under this assumption

$$Pr_Q[I(t) = 1|\mathcal{M}_t] = Pr_P[I(t) = 1|\mathcal{M}_t] = S(t,x)$$

but, for $u < t$, $Pr_Q[I(t) = 1|\mathcal{M}_u]$ and $Pr_P[I(t) = 1|\mathcal{M}_u]$ are not necessarily equal.

With an illiquid market, it is, nevertheless, plausible that insurers might be prepared to pay a premium to hedge their exposure to non-systematic mortality risk, in the same way that individuals are prepared to pay insurers a premium above the actuarially fair (i.e. P) price of an annuity in order to insure against their personal longevity risk. Under these circumstances, we need to define a risk-neutral force of mortality $\tilde{\mu}_Q(t,x)$ and the corresponding risk-neutral survivor index

$$\tilde{S}(t,x) = \exp\left[-\int_0^t \tilde{\mu}_Q(u, x+u)du\right].$$

Then $Pr_Q[I(t) = 1|\mathcal{M}_t] = \tilde{S}(t,x)$. The concept of a non-zero market price of risk for non-systematic mortality risk is explored further by Biffis *et al.* (2006).

Similarly, most modelling of stochastic mortality has assumed that the dynamics of the force of mortality curve are independent of the evolution of the term structure of interest rates. Under these circumstances, the value at time t of a pure mortality-linked cashflow X at a fixed time T can be written as the product of the zero-coupon bond price $P(t, T)$ and the risk-neutral expected value of X given the information available at time t.

2.6. Zero-coupon fixed-income and longevity bonds

In what follows, we wish to consider the value of mortality-linked cashflows. To do this, we need to introduce two types of financial asset: zero-coupon fixed-income bonds, and zero-coupon longevity bonds.

The time T-maturity zero-coupon bond pays a fixed amount of 1 unit at time T. The price at time t of this bond is denoted by $P(t, T)$. We will assume that underlying interest rates are stochastic, and the reader is referred to Cairns (2004) or Brigo & Mercurio (2001) for textbook accounts of interest-rate models. Also relevant here is a cash account $C(t)$ that constantly reinvests at the (stochastic) instantaneous risk-free rate of interest, $r(t)$, so that $C(t) = \exp[\int_0^t r(u)du]$.

Zero-coupon longevity bonds (or zero-coupon survivor bonds, as Blake & Burrows (2001) originally called them) can take two forms. Each type is characterized by a maturity date, T, and a reference cohort aged x at time 0, and we will refer to this as the (T, x)-longevity bond. Type A pays $S(T, x)$ at time T and the market value at time t of this bond will be denoted by $B_S(t, T, x)$. Type B pays $C(T)S(T, x)$ at time T, and the market value at time t of this bond will be denoted by $B_{CS}(t, T, x)$.

Throughout the rest of this paper, we will, for simplicity of exposition, assume that the term structure of interest rates is independent of the term structure of mortality, and that the market price of risk for non-systematic mortality risk is zero. It follows that (see, for example, Cairns *et al.* (2006a))

$$B_S(t, T, x) = \frac{P(t, T)B_{CS}(t, T, x)}{C(t)}.$$

Prices are linked to an artificial risk-neutral pricing measure Q (see above) as follows (see Cairns *et al.* (2006a))

$$B_{CS}(t, T, x) = E_Q\left[\frac{C(t)}{C(T)}C(T)S(T, x)\middle|\mathcal{H}_t\right] = C(t)E_Q[S(T, x)|\mathcal{M}_t]. \tag{2}$$

Here, $\mathcal{H}_t$ represents the augmented filtration that includes information about both mortality and interest rates up to time t.

Remark (Fundamental Theorem of Asset Pricing) *If there exists such a risk-neutral measure Q, and if prices are calculated according to Eq. (2), then the market will be arbitrage free. This includes markets that are incomplete or illiquid.*

It follows that $B_{CS}(t, T, x) = C(t)S(t, x)p_Q(t, T, x)$, where $p_Q(t, T, x)$ is the risk-neutral spot survival probability, $E_Q[S(T, x)/S(t, x)|\mathcal{M}_t]$. This represents the risk-adjusted probability, based on the information available at time t, that an individual aged $x+t$ at t will survive until age $x+T$.

2.6.1. Risk-neutral spot and forward survival probabilities. If the spot market in type-B bonds is sufficiently liquid for a range of maturities, then we can take as given the bond prices, $B_{CS}(t, T, x)$, at time t. This then defines the risk-neutral spot survival probabilities

$$p_Q(t,T,x)=\frac{B_{CS}(t,T,x)}{C(t)S(t,x)}$$

from which we can define the risk-neutral forward survival probabilities

$$\begin{aligned} &p_Q(t,T_0,T_1,x)=p_Q(t,T_1,x)/p_Q(t,T_0,x) \quad \text{for } t\le T_0\\ \text{and}\quad &p_Q(t,T_0,T_1,x)=\frac{S(t,x)}{S(T_0,x)}p_Q(t,T_1,x) \quad \text{for } t>T_0. \end{aligned} \tag{3}$$

2.6.2. The forward mortality surface. In the traditional mortality setting, with no mortality improvements, the force of mortality can be defined as $\mu_{x+t}=-\partial\log {}_tp_x/\partial t$. In the current context, this can be developed as a concept for individual cohorts. We assume for a given cohort aged x at time 0 that the spot survival probabilities, $p_Q(t, T, x)$ are known at time t for all $T>0$ (not just integer values of T). The forward force of mortality surface (or, more simply, forward mortality surface) is defined as (see Dahl (2004))

$$\tilde{\mu}(t,T,x+T)=-\frac{\partial}{\partial T}\log\ p_Q(t,T,x). \tag{4}$$

At a given time t, this defines a two-dimensional surface in age, x, and maturity, T. It can be regarded as a central estimate, based on information available at time t, of the force of mortality at the future time T (that is, $T-t$ years ahead) for individuals aged $x+T$ at time T.

3. Stochastic mortality models: introductory remarks

In Sections 3.1–5, we will review the full range of approaches that can be taken to modelling future randomness in mortality rates.

At least in theory (published modelling work has still to fill some gaps), there is a wide range of extrapolative approaches that can be taken to model mortality. Some models are framed in discrete time, others in continuous time. In discrete time, most research has focused on the statistical analysis and projection of annual mortality data using what will be described below as short-rate models. However, some work has also been done using P-splines and discrete-time market models. In continuous time, most research has also focused on short-rate models including affine mortality models. But other approaches analogous to forward-rate models and market models in interest-rate modelling have also been proposed or discussed.

3.1. Model selection criteria

Once a model has been developed and parameters have been estimated or calibrated, it is important to consider whether it is a good model or not. This requires a checklist of criteria against which a model can be assessed, along the lines proposed by Cairns *et al.* (2006a), Cairns *et al.* (2007) and Cairns *et al.* (2008):

- Mortality rates should be positive.
- The model should be consistent with historical data.
- Long-term dynamics under the model should be biologically reasonable.
- Parameter estimates should be robust relative to the period of data and range of ages employed.
- Model forecasts should be robust relative to the period of data and range of ages employed.
- Forecast levels of uncertainty and central trajectories should be plausible and consistent with historical trends and variability in mortality data.
- The model should be straightforward to implement using analytical methods or fast numerical algorithms.
- The model should be relatively parsimonious.
- It should be possible to use the model to generate sample paths and calculate prediction intervals.
- The structure of the model should make it possible to incorporate parameter uncertainty in simulations.
- At least for some countries, the model should incorporate a stochastic cohort effect.
- The model should have a non-trivial correlation structure.

Some of these points require further comment to bring out their relevance.

3.2. Consistency with historical data

At a minimum, a good model should be consistent with historical patterns of mortality. If this is not the case, much greater doubt must be placed on the validity of any forecasts produced by the model.

More formal statistical approaches have been used by, for example, Brouhns *et al.* (2002) and Czado *et al.* (2005), using full likelihood methods and Markov Chain Monte Carlo (MCMC) methods. Additionally, Cairns *et al.* (2007) carried out a detailed comparison of different models based on maximum likelihoods, using criteria that penalize over-parameterised models. They showed that statistically significant improvements on the Lee & Carter (1992) and Cairns *et al.* (2006b) models can be achieved by adding extra period and cohort effects. They also demonstrated that the fundamental assumption under these simple models that the two-dimensional array of standardized residuals be independent and identically distributed is violated, thereby calling into question their consistency with historical data. Further historical analysis has been performed by Dowd *et al.* (2008a,b) who use a variety of backtesting procedures to evaluate out-of-sample performance of a range of models.

3.3. Biological reasonableness

Cairns *et al.* (2006a) introduce the concept of biological reasonableness, drawing on the concept of economic reasonableness from interest-rate modelling. Different modellers might have their own idea about what constitutes a biologically reasonable model, but we offer the following examples of what might be considered biologically unreasonable:

- Period mortality tables have historically exhibited increasing rates of mortality with age at higher ages. A forecasting model that gives rise to the possibility of period mortality tables that have mortality rates falling with age *might* be considered biologically unreasonable.
- Short-term mean reversion might be considered to be biologically reasonable due to annual environmental variation. Long-run mean reversion around a deterministic trend *might*, on the other hand, be considered biologically unreasonable. In the long term, mortality improvements will, amongst other reasons, be the result of medical advances, such as a cure for cancer. It is very difficult to predict when such advances will happen or what the impact of a new treatment might be. A deterministic mean-reversion level would suggest that we do know *what* advances are going to happen and *what* their impact will be: we just do not know *when* they will happen.

3.4. Robustness of parameter estimates and forecasts

When we fit a model, we need to specify what range of ages and past years we will use to estimate parameters: for example, ages 60–89, and years 1960–2005. Ultimately, we wish to have confidence in the forecasts that are produced by a model. With most models, we find that a change in the range of years or of ages in the historical dataset results in a relatively modest revision of parameter estimates, consistent with the statistical variation in the data. Such models might be described as being robust.

For other models, however, a change in the age or calendar-year range sometimes results in (a) a set of qualitatively different parameter estimates and (b) substantially different forecasts of future mortality. Such models are not robust and they cannot be relied upon to produce consistently reasonable forecasts.

3.5. Plausibility of forecasts

The plausibility of forecasts is again a rather subjective issue, discussed by Cairns *et al.* (2008). In general, one cannot normally make a definitive statement that a set of forecasts look reasonable. However, Cairns *et al.* (2008) do provide examples of models that provide a statistically good fit to the historical data, but then produce quite implausible projections of mortality. Examples of implausible forecasts include: a sudden and dramatic change in the central trend in mortality rates at certain ages; and prediction intervals for mortality that are either extremely wide or extremely narrow.

3.6. Implementation

There is little point in having a great model if it requires excessive amounts of computing time to calculate important quantities of interest. If this happens, then a compromise needs to be reached, ideally without sacrificing too much in terms of statistical goodness of fit.

3.7. Parsimony

We should avoid models that are excessively parameterized. This can be addressed by using, for example, the Bayes Information Criterion (BIC), to choose between models. This ensures that extra parameters are only included when there is a significant improvement in fit.

3.8. Sample paths and prediction intervals

Most models (except for P-splines models) generate sample paths and therefore allow an assessment of the uncertainty in future mortality-linked cashflows, and the pricing of these cashflows. Pricing might require a change of measure and this requires a fully specified stochastic model with sample paths.

3.9. Parameter uncertainty

We have a limited amount of data that we can use to estimate model parameters, and so it follows that these parameters will be subject to estimation error. It is important to be able to incorporate parameter uncertainty, if we wish, into simulations in order to assess the impact of this estimation error. Cairns *et al.* (2006b) and, for example, Dowd *et al.* (2007) investigated this issue in the Cairns–Blake–Dowd two-factor model and found that the inclusion of parameter uncertainty has a significant impact on forecast levels of uncertainty in mortality rates and expected future lifetimes, especially at longer time horizons. We do not discuss the issue of parameter uncertainty further in this paper, but note that this, along with model risk, are important issues that require substantial further research.

3.10. Cohort effect

As remarked earlier, some countries including England and Wales, have mortality rates that appear to be determined not just by age and period effects but also by year-of-birth effects. Cairns *et al.* (2007) demonstrated that the inclusion of a cohort effect provided a statistically significantly better fit. We would expect that such effects will persist into the future and that forecasts will, therefore, be improved by the inclusion of a cohort effect.

3.11. Correlation term structure

As also remarked earlier in this paper and elsewhere, rates of improvement at different ages have been different both over long periods of time and also from one year to the next. In other words, improvements at different ages might be correlated, but they are not perfectly correlated. The use of a model that assumes perfect correlation might cause problems for two reasons. First, aggregate levels of uncertainty at the portfolio level might be overstated (since it is assumed that there are no diversification benefits across ages). Second, perfect correlation would incorrectly suggest that derivative instruments linked to mortality at one age could be used to hedge perfectly mortality improvements at a

different age. Ignoring this problem might result in a degree of hidden basis risk if hedging strategies are adopted that rely on perfect correlation.

4. Discrete-time models

National mortality data are generally published on an annual basis and by individual year of age, and this leads naturally to the development of discrete-time models. Typically, the data are presented in the form of crude death rates

$$m_c(t,x) = \frac{D(t,x)}{E(t,x)} = \frac{\#\text{deaths during calendar year } t \text{ aged } x \text{ last birthday}}{\text{average population during calendar year } t \text{ aged } x \text{ last birthday}}.$$

The average population (the exposure) will normally be an approximation, either based on a mid-year population estimate or on estimates of the population at the start and end of each year. Cairns *et al.* (2007) found, for example, that US exposures data tended to be rather less reliable than England and Wales data, leading to less reliable parameter estimates for stochastic mortality models.[2] The numbers of deaths are usually felt to be known with greater accuracy, although the accuracy might depend on local laws concerning the reporting of deaths.

Some authors have chosen to model the death rates directly, while others choose to model mortality rates, $q(t, x)$, the underlying probability that an individual aged exactly x at time t will survive until time $t+1$.

The two are linked, typically, by one of two approximations: $q(t,x) = 1 - \exp[-m(t,x)]$; or $q(t,x) = m(t,x)/(1 + 0.5m(t,x))$.

4.1. The Lee–Carter model

The availability of annual data subdivided into integer ages means that it is relatively straightforward to use rigorous statistical methods to fit discrete-time models to the data. This has tended to produce models that are straightforward to simulate in discrete time but which do not lead to analytical formulae for, for example, spot survival probabilities.

The earliest model and still the most popular is the Lee & Carter (1992) model

$$m(t,x) = \exp\left[\beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}\right]$$

where the $\beta_x^{(k)}$ are age effects and $\kappa_t^{(2)}$ is a random period effect. Under this model, $\beta_x^{(1)}$ represents an average log mortality rate over time at age x, while $\beta_x^{(2)}$ represents the improvement rate at age x. The period effect, $\kappa_t^{(2)}$, is often modelled as a random-walk process (e.g., Lee & Carter (1992)) or as an ARIMA process (e.g., CMI (2007)).

A variety of approaches have been taken to the estimation of parameters. The early approach of Lee and Carter has been largely replaced by more formal statistical methods

[2] To date, we are unaware of any studies that have explicitly attempted to model the exposures as unobserved variables.

(Brouhns *et al.* (2002), Czado *et al.* (2005) and Delwarde *et al.* (2007)) with the primary focus on goodness of fit over all of the data. A rather different approach was taken by Lee & Miller (2001) who took the view that a greater emphasis ought to be placed on goodness of fit in the final year in the dataset. They observed that the purpose of modelling is normally to project mortality rates. However, the usual statistical procedures aim to fit the historical data well over all past years. This means that the final year of historical data (sometimes referred to as the stepping-off year) might have a relatively poor fit. (Note also, that this problem tends to be worse if more calendar years of data are used.) As an example, Figure 3 shows estimated underlying mortality rates in 2004 for England and Wales mortality data (1961–2004, and ages 60–89) using the Lee–Carter model. We can see that the fitted curve systematically underestimates the death rate at lower ages. Based on an out-of-sample analysis, Lee & Miller (2001) observed that systematic bias of this type in the stepping-off year persists and it results in biased predictions of key outputs such as mortality rates at specific ages or of period life expectancy. To avoid this problem, they suggested that $\beta_x^{(1)}$ should be calibrated to log death rates in the stepping-off year, t_0 (i.e., $\beta_x^{(1)} = \log m_c(t_0, x)$) in combination with $\kappa_{t_0}^{(2)} = 0$.

Elsewhere in this paper, we will draw on interest-rate modelling analogies. In the present context, we can liken the Lee–Miller method of calibrating rates against the stepping-off-year crude death rates to the Hull & White (1990) extension of the Vasicek (1977) interest-rate model. The Vasicek model assumes a constant mean reversion level for interest rates, whereas the Hull and White model assumes a deterministic but time-varying mean-reversion level that is calibrated in a way that ensures that theoretical zero-coupon bond prices match observed prices at time zero.

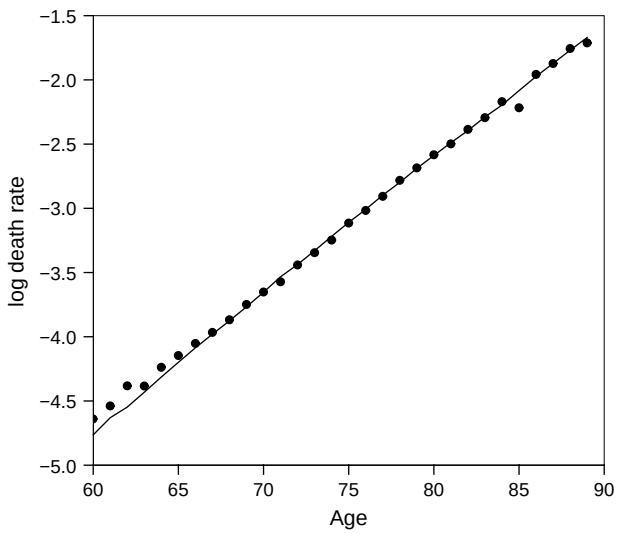

Figure 3. England and Wales males: log death rates in 2004 for ages 60–89. Dots: crude death rates. Solid line: Lee and Carter fitted curve of the underlying death rate based on data from 1961 to 2004 and ages 60–89.

Other authors have adapted the methodology in an attempt to improve the model for $\kappa_t^{(2)}$. The motivation for this (see, for example, Booth *et al.* (2002)) is that the drift of the random walk appears to change from time to time. De Jong & Tickle (2006) tackled this by modelling the drift itself as a latent random process. On the other hand, Booth *et al.* (2002) attempted to estimate the optimal estimation period for the random-walk model with constant drift. However, Booth *et al.*'s method might lead to systematic under-estimation of the true level of volatility in $\kappa_t^{(2)}$.

The Lee–Carter model serves a useful pedagogical purpose, but the model has drawbacks:

- It is a one-factor model, resulting in mortality improvements at all ages being perfectly correlated.
- The $\beta_x^{(2)}$ age effect is normally measured as the average improvement rate at age x, but $\beta_x^{(2)}$ has a second purpose in setting the level of uncertainty in future death rates at age x: $Var[\log m(T,x)|\mathcal{M}_t] = \beta_x^{(2)^2} Var[\kappa_T^{(2)}|\kappa_t^{(2)}]$. This means that uncertainty in future mortality rates cannot be decoupled from improvement rates. Historically, improvement rates have been lower at high ages. This, in turn, means that predicted uncertainty in future death rates will be significantly lower at high ages. However, this prediction does not conform with the high level of variability in mortality improvements at high ages observed in historical data, undermining the validity of the model.
- Use of the basic version of the Lee–Carter model can result in a lack of smoothness in the estimated age effect, $\beta_x^{(2)}$. Suppose that $\hat{\beta}_x^{(1)}$ and $\hat{\beta}_x^{(2)}$ have been estimated using data up to, say, 1980. We can then carry out an out-of-sample experiment by taking $\hat{\beta}_x^{(1)}$ and $\hat{\beta}_x^{(2)}$ as given and estimating $\kappa_t^{(2)}$ for each future year. We next calculate the standardized residuals for the out-of-sample period, $\varepsilon(t,x) = (D(t,x) - m(t,x)E(t,x))/\sqrt{m(t,x)E(t,x)}$. If the model and estimation method are correct, then the $\varepsilon(t,x)$ should be independent and identically distributed and approximately standard normal. Preliminary experiments indicate that if there is a lack of smoothness in $\hat{\beta}_x^{(2)}$, then there is a clear inverse relationship between the $\hat{\beta}_x^{(2)}$, for each x, and the observed out-of-sample variance of the $\varepsilon(t,x)$. This observation violates the i.i.d. standard normality assumption.

Delwarde *et al.* (2007) tackle this problem by applying P-splines (see below) to the age effects.

- The problem of bias in forecasts identified by Lee & Miller (2001) and their proposed solution arises fundamentally because the Lee–Carter model does not fit the data well. At a visual level, this is borne out by a plot of standardized residuals. The statistical approach proposed by Brouhns *et al.* (2002) assumes that deaths, $D(t, x)$ are independent Poisson random variables with mean and variance both equal to $m(t,x)E(t,x)$ where $m(t,x) = \exp[\beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}]$. The standardized residuals are

$$\varepsilon(t,x) = (D(t,x) - m(t,x)E(t,x))/\sqrt{m(t,x)E(t,x))}.$$

Standardised residuals for England and Wales males are plotted in Figure 4. If the residuals were independent then there should be no evidence of clustering. However,

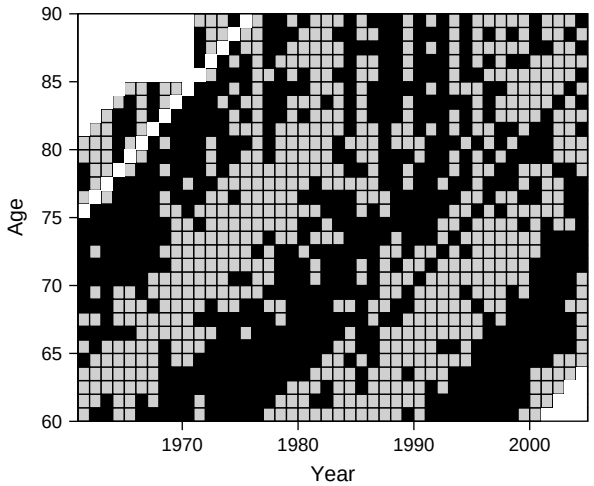

Figure 4. England and Wales males: standardized residuals, $\varepsilon(t, x)$, for the Lee–Carter model fitted to ages 60–89 and years 1961–2004. Black cells indicate a negative residual; grey cells indicate a positive residual; white cells were excluded from the analysis. (Source: Cairns *et al.* (2007))

Figure 4 shows clear evidence of clustering that violate this assumption, especially along the diagonals (the so-called cohort effect identified by Willets (2004)).

The Lee and Miller adjustment is therefore simply a 'quick fix' that improves the short-term problem of bias in forecasts. However, if the 'usual' method of estimation (for example, Brouhns *et al.* (2002)) has poor in-sample statistical properties, then the Lee–Miller calibration of the model will have the same problems with the reliability of long-term forecasts of mortality improvements.

4.2. Multifactor age–period models

A small number of multifactor models have appeared in recent years. As examples, Renshaw & Haberman (2003) propose the model

$$\log\ m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\kappa_t^{(3)}$$

where $\kappa_t^{(2)}$ and $\kappa_t^{(3)}$ are dependent period effects (for example, a bivariate random walk).

Cairns *et al.* (2006b) focus on higher ages (60–89) and used the relatively simple pattern of mortality at these ages to fit a more parsimonious model based on the logistic transform of the mortality rate rather than the log of the death rate:

$$\text{logit}\ \ q(t,x) = \log\frac{q(t,x)}{1\ -\ q(t,x)} = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x})$$

where $(\kappa_t^{(1)}, \kappa_t^{(2)})$ is assumed to be a bivariate random walk with drift. Their analysis also includes a detailed account of how parameter uncertainty can be included in simulations using Bayesian methods.

Both models offer significant qualitative advantages over the Lee–Carter model. However, both still fail to tackle the cohort effect.

4.3. The Renshaw–Haberman cohort model

Renshaw & Haberman (2006) proposed one of the first stochastic models for population mortality to incorporate a cohort effect (see also Osmond (1985) and Jacobsen *et al.* (2002)):

$$\log\ m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}$$

where $\kappa_t^{(2)}$ is a random period effect and $\gamma_{t-x}^{(3)}$ is a random cohort effect that is a function of the (approximate) year of birth, $(t-x)$.

In their analysis of England and Wales males data, Renshaw and Haberman found that there was a significant improvement over the Lee–Carter model (see, also, Cairns *et al.* (2007)). The most noticeable improvement was that an analysis of the standardized residuals revealed very little dependence on the year of birth, in contrast with the Lee–Carter model (see Figure 5). The upper plot (Lee–Carter) shows significant dependence on the year of birth, particularly between 1925 and 1935. The cohort model (lower plot) shows no such dependence.

Unfortunately, the Renshaw–Haberman model turns out to suffer from a lack of robustness. CMI (2007) found that a change in the range of ages used to fit the model might result in a qualitatively different set of parameter estimates. Cairns *et al.* (2007) and Cairns *et al.* (2008) found the same when they changed the range of years used to fit the model. This lack of robustness is thought to be linked to the shape of the likelihood function. For robust models, the likelihood function probably has a unique maximum which remains broadly unchanged if the range of years or ages is changed. For models that lack robustness, the likelihood function possibly has more than one maximum. So when we change the age or year range, the optimiser will periodically jump from one local maximum to another with qualitatively quite different characteristics.

Cairns *et al.* (2008) note a further problem with the Renshaw–Haberman model. The fitted cohort effect, $\gamma_{t-x}^{(3)}$, appears to have a deterministic linear, or possibly quadratic, trend in the year of birth. This suggests that the age–cohort effect is being used, inadvertently, to compensate for the lack of a second age–period effect, as well as trying to capture the cohort effect in the data. This suggests that an improvement on the model might be to combine the second age–period effect in Renshaw & Haberman (2003) with a simpler cohort effect. This might result in a better fit, although it might not deal with the problem of robustness.

4.4. The Cairns–Blake–Dowd model with a cohort effect

Given the problems with the preceding models, a range of alternatives have been explored with the aim of finding a model that incorporates a parsimonious, multifactor age–period structure with a cohort effect. Out of several models analysed, the following generalization

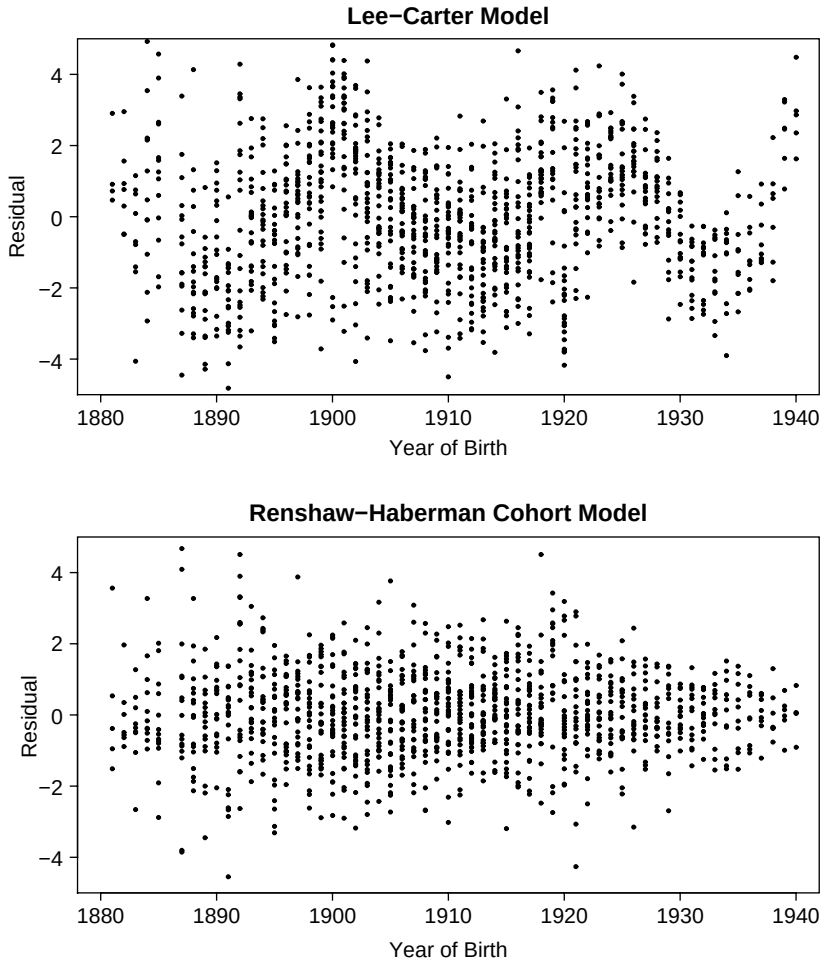

Figure 5. England and Wales, males: standardized residuals, $\varepsilon(t, x)$, for the Lee–Carter and Renshaw–Haberman cohort models plotted against year of birth, $t-x$. Both models are fitted to ages 60–89 and years 1961–2004.

of the Cairns–Blake–Dowd two-factor model (Cairns *et al.* (2006b)) has been found to produce good results (see, also, Cairns *et al.* (2007)):

$$\text{logit}\ \ q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \kappa_t^{(3)}\left((x-\bar{x})^2 - \sigma_x^2\right) + \gamma_{t-x}^{(4)}$$

where $\bar{x} = (x_u - x_l + 1)^{-1}\Sigma_{x=x_l}^{x_u} x$ is the mean in the range of ages (x_l to x_u) to be fitted, and $\sigma_x^2 = (x_u - x_l + 1)^{-1}\Sigma_{x=x_l}^{x_u}(x-\bar{x})^2$ is the corresponding variance.

Compared with the original model of Cairns *et al.* (2006b), there are two additional components. First, there is an additional age–period effect, $\kappa_t^{(3)}\left((x-\bar{x})^2 - \sigma_x^2\right)$, that is quadratic in age. For both England and Wales and US males mortality data, this additional term was found to provide a statistically significant improvement, although it is

considerably less important than the first two age–period effects. Second, we have introduced a cohort effect, $\gamma^{(4)}_{t-x}$, that is, a function of the approximate year of birth $t-x$. An alternative model that sets $\kappa^{(3)}_t$ to zero and replaces $\gamma^{(4)}_{t-x}$ with a more complex age–cohort factor, $(x_c - x)\gamma^{(4)}_{t-x}$ was also considered with mixed results (see Cairns *et al.* (2007) and Cairns *et al.* (2008)).

Parameter estimates for this model for England and Wales males aged 60–89 are plotted in Figure 6. From this we can make the following observations:

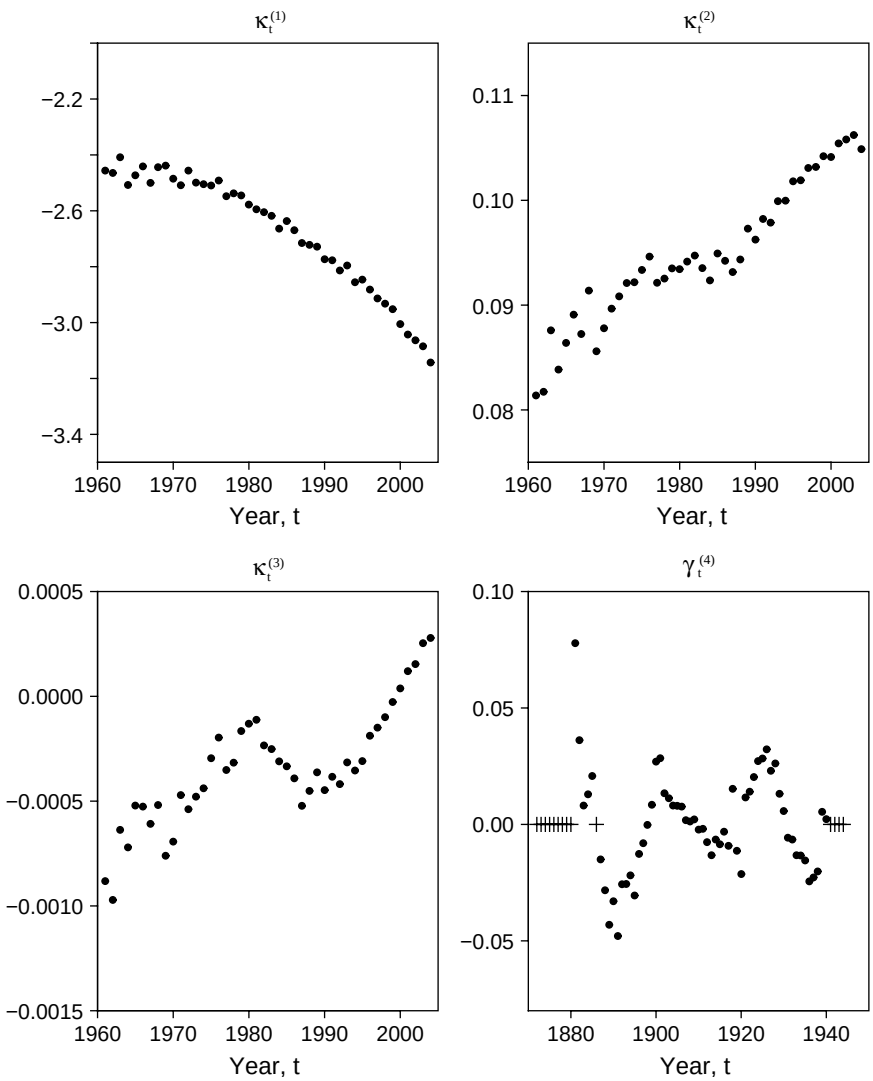

Figure 6. England and Wales data: parameter estimates for the three-factor Cairns–Blake–Dowd cohort model. Crosses in the bottom right plot correspond to excluded cohorts.

- $\kappa_t^{(1)}$ (which can be interpreted as the 'level' of mortality) has a downwards trend, reflecting generally improving mortality rates over time.
- $\kappa_t^{(2)}$ (the 'slope' coefficient) has a gradual upwards drift, reflecting the fact that, historically, mortality at high ages has improved at a slower rate than at younger ages.
- $\kappa_t^{(3)}$ (the 'curvature' coefficient) is more erratic, but has increased between 1961 and 2004.
- $\gamma_c^{(4)}$, where $c = t - x$, fluctuates around zero with no systematic trend or curvature.

For forecasting, we need to develop models for the future dynamics of the period and cohort effects. A simple starting point is to use a three-dimensional random-walk model for, $\kappa_t^{(1)}$ $\kappa_t^{(2)}$, and $\kappa_t^{(3)}$ (see, for example, Cairns *et al.* (2008)). However, there are potential dangers with this approach:

- Based on the information in Figure 6, the random-walk model for $\kappa_t^{(3)}$ is likely to have a positive drift. As $\kappa_t^{(3)}$ becomes more and more positive, the curvature of logit $q(t, x)$ would become more and more positive over time. We therefore need to question whether or not this increasing curvature could result in biologically unreasonable mortality curves. Possible compromises are to fit a random-walk model to $\kappa_t^{(3)}$ with zero drift, or to fit a process (e.g., AR(1)) that is mean reverting to some possibly non-zero level.
- The upward trend for $\kappa_t^{(2)}$ is more pronounced than it is for $\kappa_t^{(3)}$ and similar comments might apply if we were to fit a random-walk model with positive drift to $\kappa_t^{(2)}$. Specifically, this would result in mortality curves at high ages becoming gradually steeper over time, and we have to question whether or not this is biologically plausible. So, again, a model with no drift or even mean reversion might be biologically more reasonable, despite the observed strong upward trend in $\kappa_t^{(3)}$.

Although not obvious, it is by design (see the Appendix) that the cohort effect, $\gamma_c^{(4)}$, looks like a process that is mean reverting to zero. This means that it is natural to fit an AR(1) or some other mean-reverting process to $\gamma_c^{(4)}$. Making this a mean-reverting process, prevents $\gamma_{t-x}^{(4)}$ from being decomposed into age–period effects. In some other models that have been considered (see Cairns *et al.* (2007)), the cohort effect appears to have a linear drift. But linear drift in the cohort effect can often be transformed into additional age–period effects. This raises the possibility that such models could be improved by incorporating additional age–period effects and perhaps simplifying the form of the cohort effect.

4.5. P-splines

An approach that has become popular in the UK is the use of penalized splines (P-splines) (Currie *et al.* (2004) and CMI (2006)). A typical model takes the form

$$\log\ m(t,x) = \sum_{i,j}\ \theta_{ij} B_{ij}(t,x)$$

where the $B_{ij}(t, x)$ are pre-specified basis functions with regularly spaced knots, and the θ_{ij} are parameters to be estimated. It is well known that the use of splines can lead to functions that are over-fitted, resulting in fitted mortality surfaces that are unreasonably lumpy. P-splines avoid this problem by penalizing roughness in the θ_{ij} (for example, using linear or quadratic penalties). This approach has proven to be very effective at producing globally a good fit (CMI (2006)). However, excessive smoothing in the period dimension can lead to systematic over- or underestimation of mortality rates, as the model fails to capture what may be genuine environmental fluctations from one year to the next that affect all ages in the same direction (Cairns *et al.* (2007)). To avoid this problem, Currie (personal communication) is developing a variant that adds annual shocks to an underlying mortality surface that is smooth in both the age and period dimensions.

4.6. Philosophical issues

The last remarks draw attention to some philosophical differences between approaches. Specifically, to what extent should the fitted mortality surface be smooth?

A recent extension of the P-splines approach developed by Kirkby & Currie (2007) suggests that underlying mortality rates (the signal) are smooth, but we observe a signal plus noise where the noise is a systematic effect across all ages that is specific to each calendar year and which reflects 'local' conditions. The other models that have been discussed so far assume that we observe the signal with no systematic noise (the independent Poisson assumption about mortality rates). Amongst these other models, there are varying assumptions about smoothness in individual dimensions. The Lee–Carter and Renshaw–Haberman models do not assume smoothness in the age or cohort effects. In contrast, the Cairns–Blake–Dowd cohort model outlined in Section 4.4 assumes smoothness in the age effect, but not in the cohort effect.

With the limited amount of data at our disposal, it remains a matter for debate as to how much smoothness should be assumed.

4.7. Discrete-time market models

In the field of interest-rate modelling, market models are ones that model the rates that are directly traded in the market such as LIBOR rates or swap rates. There are a number of equivalent approaches that can be taken in the stochastic modelling of mortality rates. In this section, we will consider a hypothetical market in zero-coupon longevity bonds with a full range of ages and terms to maturity. The price at time t, as remarked earlier, of the (T, x)-bond that pays $C(T)S(T, x)$ at time T is denoted by $B_{CS}(t, T, x)$.

For the market to be arbitrage free, we require the existence of a risk-neutral measure, Q, under which the discounted asset price, $B_{CS}(t, T, x)/C(t) = E_Q[S(T, x)|\mathcal{M}_t]$, is a martingale. Using the terminology of forward survival probabilities (Eqs. (1) and (3)), we have $p_Q(t, T_0, T_1, x) = B_{CS}(t, T_1, x)/B_{CS}(t, T_0, x)$. The martingale property of discounted asset prices means that

$$p_Q(t, t, T, x) = E_Q\left[p_Q(t+1, t, T, x)|\mathcal{M}_t\right] \tag{5}$$

for all $t < T$ and for all x. (Recall that (Eq. (3)) $p_Q(t+1, t, T, x) = \{S(t+1, x)/S(t, x)\} \times p_Q(t+1, t+1, T, x)$.)

Market models simultaneously model the prices of all of the type-B (T, x)-longevity bonds. This approach was first adopted in the Olivier–Smith model (see Olivier & Jeffery (2004) and Smith (2005)) which assumes for all $t \leq T$ and for all x:

$$p_Q(t+1, T, T+1, x) = p_Q(t, T, T+1, x)^{b(t+1,\, T,\, T+1,\, x)G(t+1)}$$

where $G(1)$, $G(2)$, ... is a sequence of independent and identically distributed Gamma random variables with both shape and scaling parameters equal to some constant α. The function $b(t+1, T, T, x)$ is measurable at time t and is a normalizing constant that ensures that Eq. (5) is satisfied. This suffers the drawback that it is a one-factor model that results in (a) lack of control over variances of mortality rates at individual ages and (b) perfect correlation between mortality improvements at different ages.

The use of a Gamma random variable has two advantages over possible alternatives. It makes the calculations analytically tractable, and it ensures that risk-neutral forward survival probabilities remain between 0 and 1.

Cairns (2007) proposes a generalization of the Olivier–Smith model that moves away from dependence on a single source of risk and allows for full control over the variances and correlations:

$$p_Q(t+1, t, T, x) = p_Q(t, t, T, x)e^{g(t+1,\, T,\, x)G(t+1,\, T,\, x)}. \tag{6}$$

In this model, the $g(t+1, T, x)$ are again normalizing constants that ensure the martingale property is satisfied. The $G(t+1, T, x)$ are dependent Gamma random variables that are specific to each maturity date T and to each age x. Cairns (2007) leaves as an open problem how to generate the Gamma random variables and provides a detailed discussion of why this is, potentially, a difficult challenge. However, if this problem can be overcome then the generalized model has considerable potential.

5. Continuous-time models

Even though data are typically reported in aggregate and at discrete intervals (e.g., annually), it is natural to consider the development of mortality rates in continuous time. Cairns *et al.* (2006a) identify four frameworks that can be used in the development of continuous-time models:[3]

- short-rate models – models for the one-dimensional (in x) instantaneous force of mortality $\mu(t, x)$;
- forward-rate models – models for the two-dimensional (in age x, and maturity T) forward mortality surface, $\tilde{\mu}(t, T, x+T)$ (Eq. (4));

[3] The discrete-time models described in Section 4 can all be described as short-rate models, with the exception of the market model in Section 4.7.

- market models – models for forward survival probabilities or annuity prices; and
- positive mortality models – models for the spot survival probabilities.

No models have been proposed so far under the positive mortality framework, so this approach is not discussed further in this article.

5.1. Philosophy: diffusions or jumps

Before we progress to consider specific continuous-time frameworks and models, we will digress briefly to consider what underlying stochastic processes to use. Most modelling work to date has focused on the use of diffusion processes: that is, mortality dynamics are driven by one or more underlying Brownian motions. However, there is the possibility that these Brownian motions might be replaced or supplemented by the use of jump processes. Jumps in the mortality process might occur for a variety of reasons: sudden changes in environmental conditions, or radical medical advances.

If they are included, jumps might have one of two effects. In the simplest case, they have a direct impact on the instantaneous force of mortality, $\mu(t, x)$. Examples of this might include a sudden pandemic or war. However, it is plausible that the sudden announcement at time t of, for example, a cure for cancer might not have such an immediate impact. Instead, the cure might take time to become widely used and implemented effectively. This would imply that there is no jump in the spot instantaneous force of mortality. However, the discovery might cause a jump in the improvement rate of $\mu(t, x)$. As a consequence, forward mortality rates would also jump, with larger jumps the further into the future.

Models with jumps are in their infancy (see, for example, Hainaut & Devolder (2008) and Chen & Cox (2007)), but their use will require just as much care in terms of model evaluation (see Section 3.1). Thus, for example, the inclusion of jumps will need to be motivated by sound biological reasoning, rather than their inclusion for the sake of mathematical generalization.

As a final remark, the discrete nature of mortality data inevitably makes the use of jump models versus diffusion models a matter of speculation: if we only observe the process annually then we cannot observe jumps directly: instead, we can only observe the compounded effect of jumps. Again, therefore, this places a degree of responsibility on modellers to provide a good biological justification for the inclusion of jumps.

5.2. Short-rate models

Continuous-time short-rate models have, to date, proved the most fruitful source of new models. Models are of the type:

$$d\mu(t,x) = a(t,x)dt + b(t,x)'d\tilde{W}(t)$$

where $a(t, x)$ is the drift, $b(t, x)$ is an $n \times 1$ vector of volatilities, and $\tilde{W}(t)$ is a standard n-dimensional Brownian motion under the risk-neutral measure, Q. Risk-neutral spot survival probabilities (assuming an arbitrage-free market) are then given by (see, for example, Milevsky & Promislow (2001) or Dahl (2004))

$$p_Q(t,T,x) = E_Q\left[\exp\left(-\int_t^T \mu(u,x+u)du\right)\middle|\mathcal{M}_t\right].$$

The drift and volatility processes will almost certainly depend on the current term structure of mortality to ensure that the mortality curve remains positive and retains a biologically reasonable shape. (For further discussion, see Cairns *et al.* (2006a).)

If $a(t, x)$ and $b(t, x)$ satisfy certain conditions (see Dahl (2004)), we have a closed-form expression for the spot survival probabilities:

$$p_Q(t,T,x) = \exp[A_0(t,T,x) - A_1(t,T,x)\mu(t,x+t)]. \tag{7}$$

Models where the logarithm of the spot survival probability is a linear function of the current force of mortality are referred to as affine models, and have become the most popular of the stochastic mortality models for their analytical tractability (Dahl (2004), Biffis (2005), Biffis & Millossovich (2006), Dahl & Møller (2006) and Schrager (2006)). Non-affine models have been developed by Milevsky & Promislow (2001) and Biffis *et al.* (2006).

Dahl & Møller (2006) develop a one-factor mortality model and use the concept of mortality improvement factors to relate the future force of mortality to the current mortality term structure:

$$\mu(t,x+t) = \mu(0,x+t)\xi(t,x+t)$$

where $\xi(t, x+t)$ is the improvement factor with dynamics

$$d\xi(t,y) = (\gamma(t,y) - \delta(t,y)\xi(t,y))dt + \sigma(t,y)\sqrt{\xi(t,y)}d\tilde{W}(t).$$

Thus, $\xi(t, y)$ takes the form of a time-inhomogeneous Cox *et al.* (1985) model (CIR). This model satisfies the criterion for an affine mortality term structure and so spot survival probabilities take the form (7). The deterministic functions $\gamma(t, y)$ and $\delta(t, y)$ allow for considerable flexibility in terms of the rate at which mortality is predicted to improve over time. It does require mean reversion (that is, $\delta(t, y) > 0$; see Section 3.3), but it is argued that this mean reversion can be relatively weak and, therefore, is a minor disadvantage when compared with the advantages of tractability. Dahl & Møller (2006) give examples for $\gamma(t, y)$, $\delta(t, y)$ and $\sigma(t, y)$ that take particular parametric forms that are then calibrated approximately to Danish mortality data. A more general two-factor model is proposed by Biffis & Millossovich (2006).

5.3. The forward-mortality modelling framework

This approach has been outlined by a number of authors (see, for example, Dahl (2004), Miltersen & Persson (2005), Bauer (2006), Bauer *et al.* (2008) and Cairns *et al.* (2006a)) all echoing the original work of Heath *et al.* (1992) in an interest-rate setting. The general form of the model has the following dynamics:

$$\tilde{\mu}(t,T,x+T) = \alpha(t,T,x+T)dt + \beta(t,T,x+T)'d\tilde{W}(t)$$

where $\alpha(t, T, y)$ is a scalar, $\beta(t, T, y)$ is an $n \times 1$ vector and $\tilde{W}(t)$ is a standard n-dimensional Brownian motion under the risk-neutral measure, Q. For the market to be arbitrage free we require

$$\alpha(t,T,x+T) = -V(t,T,x)'\beta(t,T,x+T)$$

where $\beta(t,T,x+T) = \partial V(t,T,x)/\partial T$.

Miltersen & Persson (2005) also consider the case where mortality and interest rates are correlated and this enriches the modelling environment given that mortality dynamics now involve the Brownian motions that drive interest rates.

5.4. Change of numeraire and market models

In the fixed-income markets, modelling in recent years has focused on the use of the LIBOR and swaps market models (Brace *et al.* (1997), Jamshidian (1997) and Miltersen *et al.* (1997)). In mortality modelling, the equivalent models are the SCOR market model (subsection 5.4.2) and the annuity market model as proposed by Cairns *et al.* (2006a).

5.4.1. Change of numeraire. The development of market models relies heavily on the potential to change between different pricing measures. Until now, we have focused on the use of the risk-neutral measure, Q, for calculating prices. Under Q, the prices of all tradeable assets discounted by the cash account, $C(t)$, are martingales. Recall, therefore, the price of the type-B (T, x)-longevity bond (Eq. (2)). The discounted asset price is

$$\tilde{B}_{CS}(t,T,x) = \frac{B_{CS}(t,T,x)}{C(t)} = E_Q[S(T,x)|\mathcal{M}_t].$$

In a continuous-time diffusion setting, the dynamics of this discounted price process can be written as

$$d\tilde{B}_{CS}(t,T,x) = \tilde{B}_{CS}(t,T,x)V(t,T,x)'d\tilde{W}(t)$$

where $V(t, T, x)$ is a previsible $n \times 1$ process and $\tilde{W}(t)$ is a standard n-dimensional Brownian motion under Q. In these expressions, we have used the cash account $C(t)$ as the numeraire. But we can use the prices of other tradeable assets (provided these prices remain positive) as alternative numeraires. For example, if we take $B_{CS}(t, \tau, x)$ as the numeraire, and define $Z(t,T,x) = B_{CS}(t,T,x)/B_{CS}(t,\tau,x)$, then

$$dZ(t,T,x) = Z(t,T,x)(V(t,T,x) - V(t,\tau,x))'(d\tilde{W}(t) - V(t,\tau,x)dt).$$

We now define $dW^{\tau,x}(t) = d\tilde{W}(t) - V(t,\tau,x)dt$ and note that this depends only on the volatility of the numeraire and not on any characteristics of the (T, x)-longevity bond. Provided that $V(t, \tau, x)$ satisfies the Novikov condition (see, for example, Karatzas & Shreve (1998)), there exists a measure $P_{\tau,x}$ equivalent to Q under which the resulting process, $W^{\tau,x}(t)$, is a standard Brownian motion. Under $P_{\tau,x}$ the $Z(t, T, x)$ are martingales for all $t < T$ and $t < \tau$: that is

$$dZ(t,T,x) = Z(t,T,x)(V(t,T,x) - V(t,\tau,x))'dW^{\tau,x}(t).$$

Why is a change of numeraire relevant? First, in interest-rate modelling, it was discovered that the pricing of interest-rate derivatives could sometimes be simplified by making a change of measure (see, for example, Cairns (2004)). Second, the LIBOR and swap market models both rely on a change of measure as a fundamental step in defining the model dynamics. Third, if we make certain assumptions about the volatility, then the use of

models that exploit a change of numeraire makes it straightforward to value key option contracts.

5.4.2. The SCOR market model. The survivor credit offered rate (SCOR; first proposed by Cairns *et al.* (2006a)) is similar to the bonus in a traditional with-profits insurance contract where the bonus is linked solely to the historical development of mortality rates. A minor difference, though, is that the SCOR represents a bonus rate that is set one year in advance of its payment. By analogy with forward LIBOR rates, the forward SCOR $s(t, T, T+1, x)$ can be linked to a financial and mortality-linked derivative with the following terms (for variants on the contract below, see Cairns *et al.* (2006a)):

- At time t, no money exchanges hands, and the value of the derivative is zero.
- At time T, the contract-holder will pay a fixed amount, K, to the counterparty, to be invested in the risk-free cash account, $C(u)$.
- At time $T+1$, the contract-holder will receive

$$K \times C(T+1)/C(T) \times (1 + s(t, T, T+1, x)) \times (1 - q(T, x+T))$$

- where $q(T, x+T)$ is the realized mortality rate between T and $T+1$ of individuals aged $x+T$ at time T: that is, $1 - q(T, x+T) = S(T+1, x)/S(T, x)$.

The rate that ensures that this contract has zero value at time t is (see Cairns *et al.* (2006a))

$$\begin{aligned} s(t, T, T+1, x) &= \frac{p_Q(t, T, x) - p_Q(t, T+1, x)}{p_Q(t, T+1, x)} \\ &= \frac{B_{CS}(t, T, x) - B_{CS}(t, T+1, x)}{B_{CS}(t, T+1, x)}. \end{aligned} \tag{8}$$

Variants of this contract can link the initial payment to financial markets up to T or, for example, to the survivor index itself (Cairns *et al.* (2006a)).

The forward SCOR (Eq. (8)) is therefore the ratio of the value of a portfolio (long in the (T, x)-longevity bond and short in the $(T+1, x)$-longevity bond) to the price of the $(T+1, x)$-longevity bond. It follows that the forward SCOR, $s(t, T, T+1, x)$ must be a martingale under the $(T+1, x)$-forward measure $P_{T+1, x}$. We can, therefore, mimic the LIBOR market model by postulating that if the volatility of $s(t, T, T+1, x)$ is deterministic, then $s(T, T, T+1, x)$ will be log-normal under $P_{T+1, x}$ (an admissible distribution, since the forward SCOR can take any value between 0 and ∞). This property then allows straightforward construction of option-type derivatives on $s(T, T, T+1, x)$ with Black–Scholes-type pricing formulae (Cairns *et al.* (2006a)).

5.4.3. Individual insurance. In an insurance context, the contract-holder is a policyholder who is aged $x+t$ at time t, with survival function $I(u) = 1$ if the policyholder is still alive at time u and 0 otherwise. The cashflows to the policyholder are then replaced by $-KI(T)$ at time T and $K\dfrac{C(T+1)}{C(T)}(1 + s(t, T, T+1, x))I(T+1)$. K might be linked to

financial and other events up to time T. The SCOR then represents a bonus rate payable to survivors, while assets belonging to policyholders who die revert to the insurer. A traditional annuity contract is, indirectly, an example of such an arrangement. Provided the market price of non-systematic mortality risk is zero, this variant of the SCOR is priced in the same way as above (see Cairns *et al.* (2006a)).

5.5. *Forward-rate and market models: advantages and disadvantages*

In certain cases, forward-rate and market models offer the possibility of analytical tractability. But the key advantage of forward-rate and market models is that, at any future date t, they automatically provide as output a two-dimensional forward mortality table ($p_Q(t, T, x)$ for all T and x). This means that it is easy to compute the payoff on a contract that is linked to the two-dimensional mortality table that is in use at time t (for example, a guaranteed annuity contract). With some exceptions (e.g., Dahl (2004)), it is not straightforward to compute such a two-dimensional table using short-rate models. In some cases (see, for example, Dowd *et al.* (2007)) this problem can be mitigated through the use of good analytical approximations for key outputs as functions of the finite number of underlying state variables.

But forward-rate and market models have their limitations. First, they require as input at time 0, a complete two-dimensional set of market spot survival probabilities. At the present time, such *market-determined* variables do not exist. (Market annuity prices do offer a partial solution, but these can be heavily distorted by expense loadings.) Instead, it will be necessary to use tables of projected survival rates using other models. Second, the longest maturity date that is input at time 0 places a limit on the length of each sample path: that is, we can only follow the paths of SCOR rates, $s(t, T, T+1, x)$, that are input at time 0, and each SCOR rate 'expires' at its respective payment date $T+1$.

6. The management of mortality risk

Recent years have seen a growing realization that mortality risk can be significant for financial institutions such as life assurers and pension plans (see, for example, O'Brien (1999)). Depending upon how such institutions have invested their assets, mortality risk might not be the largest risk they face, but it is often significant and one that cannot be ignored. Instead, modern risk-management practice requires companies to manage mortality risk as effectively as possible as part of a wider framework of enterprise risk management rather than to accept its presence as inevitable.

A range of possible responses to longevity risk is possible, some depending on the type of institution (see, for example, Blake *et al.* (2006)).

- Assurers can retain mortality risk as a legitimate business risk. This assumes that the company is able to achieve an adequate expected rate of return relative to the level of risk being carried.

- Assurers can diversify their mortality risk across product ranges, regions and socio-economic groups. An example of this includes natural hedging (see, for example, Cox & Lin (2007)) where gains on the life book will balance losses on the annuity book.
- They can enter into a variety of forms of full or partial reinsurance, in order to hedge downside mortality risk.
- Pension plans can arrange a full or partial buyout of their liabilities by a specialist insurer (e.g., Paternoster in the UK). (Small pension plans in the UK are exposed to considerable non-systematic mortality risk and often, therefore, purchase annuities from a life office for employees at the time of their retirement, thereby removing the tail mortality risk.)
- For future annuity and pension provision, non-profit contracts could be replaced by participating annuities. Such annuities might share mortality profits or losses by adjusting the amounts of pensions in payment or by linking the date of retirement to current life expectancy.
- Assurers can securitize a line of business (see, for example, Cowley & Cummins (2005)).
- Mortality risk can be managed through the use of mortality-linked securities and derivatives. This approach differs from the securitisation of a line of business as the financial securities concerned have cashflows that are purely linked to the future value of a mortality index, rather than being a complex package of business risks.

We will focus here on the last of these possibilities, mortality-linked securities and derivatives, which have recently been made available by the establishment of a new capital market called the 'life market'. Loeys *et al.* (2007, p. 6) show that for a new capital market to become established and to flourish: *'it (1) must provide effective exposure, or hedging, to a state of the world that is (2) economically important and that (3) cannot be hedged through existing market instruments, and (4) it must use a homogeneous and transparent contract to permit exchange between agents.'* They argue that a market trading mortality risk meets these criteria and that a successful market should emerge in due course.

We will now describe briefly the different types of security that could be used.

6.1. Mortality catastrophe bonds

Although this paper is primarily concerned with mortality risk, it is, nevertheless, instructive to discuss short-dated mortality catastrophe bonds, since there have been a number of successful issues of this type of bond. They are market-traded securities whose payments are linked to a mortality index. They are similar to catastrophe bonds.

The first such bond issued was the Swiss Re bond (known as Vita I) which came to market in December 2003. This was designed to securitize Swiss Re's own exposure to mortality risk. Vita I was a three-year bond (maturing on 1 January 2007) which allowed the issuer to reduce exposure to certain catastrophic mortality events: a severe outbreak of influenza, a major terrorist attack (using weapons of mass destruction) or a natural catastrophe. The mortality index was weighted by age, sex and nationality. The $400m principal was at risk if, during any single calendar year, the combined mortality index

exceeded 130% of the baseline 2002 level, and would be exhausted if the index exceeded 150%. In return for having their principal at risk, investors received quarterly coupons of three-month US LIBOR plus 135 basis points.

The success of the bond led to additional bonds being issued on much less favourable terms to investors: for example, Vita II by Swiss Re in 2005 ($362m), Vita III by Swiss Re in 2007 ($705m), Tartan by Scottish Re in 2006 ($155m) and OSIRIS by AXA in 2006 ($442m).

6.2. Mortality (or survivor) swaps

The key derivative of interest is the mortality (or survivor) swap (e.g., Dowd *et al.* (2006), see also Lin & Cox (2005)). Counterparties swap fixed series of payments in return for series of payments linked to the number of survivors in a given cohort. One example would be a swap based on 65-year-old males from England and Wales. As another example, a UK annuity provider could swap cashflows based on a UK mortality index for cashflows based on a US mortality index from a US annuity provider counterparty: this would enable both counterparties to diversify their longevity risks internationally.

The world's first publicly announced mortality swap took place in April 2007 between Swiss Re and Friends' Provident, a UK life assurer. It was a pure longevity risk transfer and was not tied to another financial instrument or transaction. The swap was based on Friends' Provident's £1.7bn book of 78,000 of pension annuity contracts written between July 2001 and December 2006. Friends' Provident retains administration of policies. Swiss Re makes payments and assumes longevity risk in exchange for an undisclosed premium. However, it is important to note that this particular swap was legally constituted as an insurance contract and was not a capital market instrument.

In a further development in December 2007, Goldman Sachs launched a monthly index called QxX.LS (www.qxx-index.com) in combination with standardized 5 and 10-year mortality swaps. The index is based on a pool of 46,290 anonymized lives over the age of 65 from a database of life-policy sellers (the so-called life-settlements market) assessed by the medical underwriter AVS.

In a thorough and groundbreaking paper, Dahl *et al.* (2008) consider a situation where a mortality (or survivor) swap is used to hedge dynamically the mortality risk in a life book. An important contribution in Dahl *et al.*'s work is the acknowledgement that the reference population underlying the swap might be different from that in the portfolio being hedged, and they carry out a detailed numerical example to illustrate the impact of the resulting basis risk. The tractability offered by the affine model they use allows a more straightforward development of hedging strategies for mitigating mortality risk (one of our criteria for a good model: see subsection 3.1). In the case of Dahl *et al.*'s (2008) work, the analytical tractability of their affine mortality model makes it possible to implement and evaluate the effectiveness of a dynamic hedging strategy. This recalls our earlier model evaluation criterion that it should be straightforward to implement a model either analytically or using efficient numerical algorithms.

6.3. Longevity (or survivor) bonds

The world's first attempt to issue a longevity (or survivor) bond was in November 2004. The European Investment Bank (EIB) offered to issue a 25-year longevity bond with an issue price of £540m and an initial coupon of £50m. The reference survivor index, $S(t)$, was based on 65-year-old males from the national population of England and Wales as produced by the UK Government Actuary's Department (GAD). The structurer/manager was BNP Paribas which assumed the longevity risk, but reinsured it through PartnerRe, based in Bermuda. The structure of the bond (see Blake *et al.* (2006), Figure 4) involves both a mortality swap and an interest-rate swap. The target group of investors was UK pension funds, but, for reasons discussed in Blake *et al.* (2006), the bond did not attract sufficient investor interest and was later withdrawn.

6.4. Longevity-linked securities

A perceived problem with the EIB longevity bond was that the reference index might not be sufficiently highly correlated with a hedger's own mortality experience (as a result of basis risk). An alternative instrument, that we will refer to as a longevity-linked security (LLS), deals, at least partly, with this problem. The concept is inspired by the design of mortgage-backed securities. The LLS is built around a special purpose vehicle (SPV). Individual hedgers on one side of the contract (for example, pension plans) arrange mortality swaps with the SPV using their own mortality experience at rates that are negotiated with the SPV manager. The swapped cashflows are then aggregated and passed on to the market. Bondholders gain if mortality is heavier than anticipated.

It might be felt that the aggregate cashflows themselves lack transparency (this did not seem to be a problem with mortgage-backed securities until the emergence of the credit crunch in 2007) in which case the SPV might link cashflows to an accepted reference index. The difference between this and the aggregated swap cashflows is a basis risk that is borne by the SPV manager.

This type of arrangement is illustrated in Figure 7. In this example, there are three hedgers, A, B, and C. Hedger A wishes to swap the risky longevity-linked cashflows $L_A(t)$ for a series of pre-determined cashflows. The agreement with the SPV manager is to swap floating $L_A(t)$ for fixed $\tilde{L}_A(t)$ for $t = 1, \ldots, T$, with the fixed leg set at a level that results in the swap initially having zero value at time 0. Similarly, hedger B swaps floating $L_B(t)$ for fixed $\tilde{L}_B(t)$, and hedger C floating $L_C(t)$ for fixed $\tilde{L}_C(t)$. The SPV itself invests in AAA-rated, fixed-interest securities of appropriate duration or uses floating rate notes plus an interest-rate swap. The LLS bondholders pay an initial premium that is used to buy the fixed-interest securities and to pay an initial commission to the manager. The bondholders in return receive coupons and, possibly, a final repayment of principal that is linked either to the hedgers' floating cashflows or to a reference index that matches as closely as possible the combined cashflows. In the latter case, any differences accrue to the SPV manager. The bondholders will not normally be hedgers themselves, so they will expect a fair premium over market fixed-interest rates in return for assuming the mortality risk.

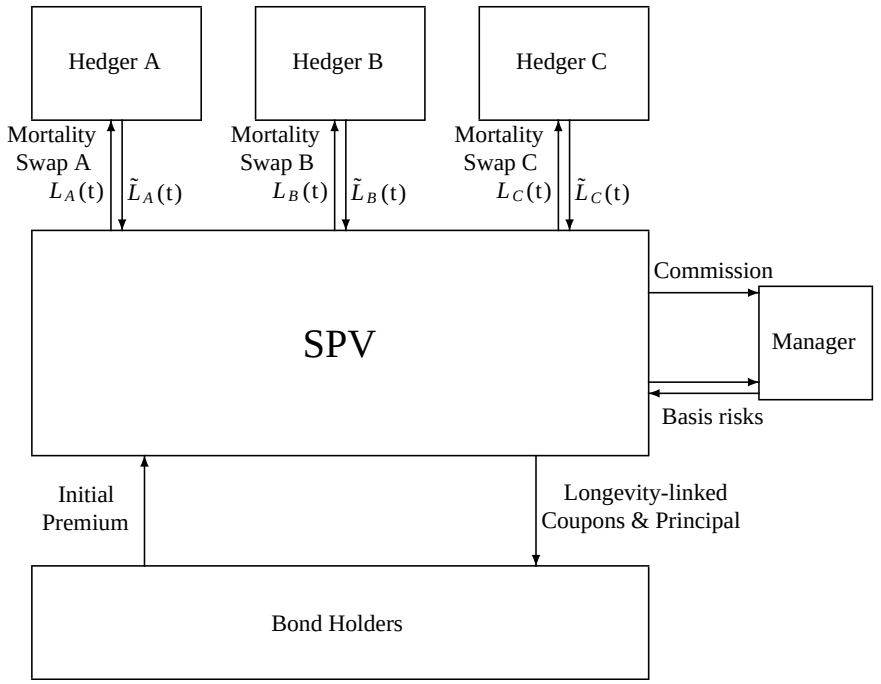

Figure 7. Cashflows under a longevity-linked security (LLS). Bondholders might receive cashflows linked to a reference index rather than $L_A(t)$, $L_B(t)$ and $L_C(t)$, in which case residual basis risk must be borne by the SPV manager.

6.5. Mortality (or q-) forwards

In July 2007, JPMorgan announced the launch of a mortality forward contract with the name 'q-forward' (Coughlan *et al.* (2007)). It is a forward contract linked to a future mortality rate, taking its name from the standard actuarial notation, 'q', for a mortality rate. The contract involves the exchange, at time $T+1$, of a realized mortality rate, $q(T, x)$, relating to a specified population on the maturity date of the contract, in return for a fixed mortality rate agreed at the beginning of the contract (Figure 8).

A series of q-forward contracts, with different ages, can be combined to hedge, approximately, a mortality swap. As an example, suppose the contract involves swapping at time t a fixed cashflow, $\hat{S}(t)$, for the realized survivor index, $S(t, x)$. The fixed leg can be

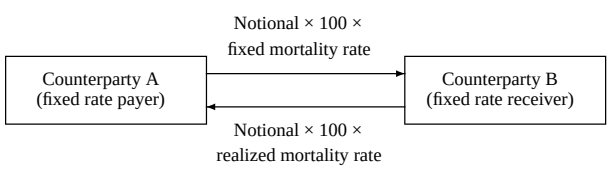

Figure 8. A q-forward arranged at time t exchanges fixed mortality, $q_F(t, T, y)$, for realized mortality, $q(T, y)$, at maturity, $T+1$, of the contract. (Source: Coughlan *et al.* (2007, Figure 1))

hedged using zero-coupon fixed-income bonds. The floating leg can be hedged approximately as follows. First, note that we can approximate the survivor index by expanding the cashflow in terms of the fixed legs of a set of q-forwards and their ultimate net payoffs:

$$\begin{aligned} S(t,x) &= (1-q(0,x)) \times (1-q(1,x+1)) \times \ldots \times (1-q(t-1,x+t-1)) \\ &= \prod_{i=0}^{t-1} (1-q_F(0,i,x+i) - \Delta(i,x+i)) \\ &\approx \prod_{i=0}^{t-1} (1-q_F(0,i,x+i)) - \sum_{i=0}^{t-1} \Delta(i,x+i) \prod_{j=0, j\neq i}^{t-1} (1-q_F(0,j,x+j)) \end{aligned}$$

where $\Delta(i,x+i) = q(i,x+i) - q_F(0,i,x+i)$ and $q_F(0,i,x+i)$ = q-forward mortality rate (the fixed rate). $\Delta(i, x+i)$ is the net payoff on the q-forward per unit at time $i+1$.

It follows that an approximate hedge (assuming interest rates are constant and equal to r per annum) for $S(t, x)$ can be achieved by holding:

- $-(1+r)^{-(t-1)} \prod_{j=0, j\neq 0}^{t-1} (1-q_F(0,j,x+j))$ units of the 1-year q-forward;
- $-(1+r)^{-(t-2)} \prod_{j=0, j\neq 1}^{t-1} (1-q_F(0,j,x+j))$ units of the 2-year q-forward;
- $\vdots$
- $-\prod_{j=0, j\neq (t-1)}^{t-1} (1-q_F(0,j,x+j))$ units of the t-year q-forward.

In calculating these hedge quantities, we take account of the fact that, for example, the payoff at time 1 on the 1-year q-forward will be rolled up to time t at the risk-free rate of interest. Hence, the required payoff at time t needs to be multiplied by the discount factor $(1+r)^{-(t-1)}$. In a stochastic interest environment, a quanto derivative would be required. A quanto derivative is one that delivers a number of units, N, of a specified asset, where N is derived from a reference index that is different from the asset being delivered. In this context, N equals $-\Delta(i,x+i) \prod_{j=0, j\neq i}^{t-1} (1-q_F(0,j,x+j))$, and we deliver, at time $i+1$, N units of the fixed-interest zero-coupon bond maturing at time t, price $P(i+1, t)$ at time $i+1$ per unit.

The discussion above covers the 'pure' version of the q-forward contract. However, some practical issues come into play. First, the hedging arguments here would need modification if, as is likely, q-forwards are subject to annual margin calls that result in the contracts being marked-to-market. Second, in order to keep the number of contracts to a manageable level, individual contracts use the average (or 'bucketed') mortality across 10 ages rather than single ages. This averaging has positive and negative effects. On the one hand, the averaging reduces the basis risk that arises from the non-systematic mortality risk that is present in crude mortality rates, even at the population level. On the other hand, it introduces some basis risk depending on the specific age-structure of the population being hedged.

6.6. Forward SCOR

For completeness, we will recall subsection 5.4.2 where we introduced the forward SCOR contract. This contract is closely linked to the q-forward contract since the payoff structure involves both the SCOR rate and the mortality rate underlying the q-forward. The two are related, additionally, by the relationships.

$$E_Q[q(T,x)|\mathcal{M}_T] = \frac{s(T,T,T+1,x-T)}{1+s(T,T,T+1,x-T)}$$
$$\Rightarrow E_Q[q(T,x)|\mathcal{M}_t] = E_Q\left[\frac{s(T,T,T+1,x-T)}{1+s(T,T,T+1,x-T)}\middle|\mathcal{M}_t\right] \quad \text{for } t<T.$$

6.7. *Mortality options, and longevity (or survivor) caps and floors*

Mortality options give payoffs that are non-linear functions of underlying variables and are attractive to (a) hedgers who might wish to protect their downside exposure, but leave any upside potential, and (b) speculators who want to trade views on volatility rather than views on the level of mortality (or related, e.g., annuity) rates.

In the case of longevity (or survivor) caps and floors, the underlying is a survivor index $S(t, x)$. Suppose $s_c(t)$ is a cap rate for exercise date t, while $s_f(t)$ is the corresponding floor rate. A caplet will pay $\max\{S(t, x) - s_c(t), 0\}$ at time t, while a floorlet will pay $\max\{s_f(t) - S(t, x), 0\}$. Series of caplets and floorlets then get packaged into caps and floors.

7. Conclusions

There have been tremendous conceptual advances in both the modelling and management of mortality risk in recent years. These advances are not unrelated. A good stochastic mortality model is a virtual prerequisite to the successful introduction of a capital market in mortality hedging instruments. Where we consider the development of other markets, such as that for interest-rate derivatives, we can note that the early days were characterized by pricing-to-model, but once the market had become sufficiently deep and liquid, the models could be dropped in favour of pricing-to-market. The same is likely to happen in the life market. But the development of a good and reliable model requires time and considerable patience: an initial analysis might suggest that a model is satisfactory, but further forensic investigation might reveal some pitfalls that need corrective work. However, this development time can repay itself handsomely through greater understanding of the underlying process driving mortality rate dynamics.

There are still considerable challenges ahead. The existing models need further refinement; the market in mortality-linked derivatives is still in its infancy; and we still need to develop further our understanding of how these contracts can be used in the most effective way to manage mortality risk.

References

Bauer, D. (2006). An arbitrage-free family of longevity bonds. Working paper, University of Ulm. Available at: www.mortalityrisk.org (accessed 11 February 2006).

Bauer, D., Boerger, M. & Russ, J. (2008). On the pricing of longevity-linked securities. Working paper, University of Ulm. Available at: www.mortalityrisk.org (accessed 20 May 2008).

Biffis, E. (2005). Affine processes for dynamic mortality and actuarial valuations. *Insurance: Mathematics and Economics* **37**, 443–468.
Biffis, E., Denuit, M. & Devolder, P. (2006). Stochastic mortality under measure changes. Pensions Institute Discussion Paper PI–0512. (April 19, 2006 version)
Biffis, E. & Millossovich, P. (2006). The fair value of guaranteed annuity options. *Scandinavian Actuarial Journal* **1**, 23–41.
Blake, D. & Burrows, W. (2001). Survivor bonds: helping to hedge mortality risk. *Journal of Risk and Insurance* **68**, 339–348.
Blake, D., Cairns, A. J. G. & Dowd, K. (2006). Living with mortality: longevity bonds and other mortality-linked securities, (with discussion). *British Actuarial Journal* **12**, 153–228.
Booth, H., Maindonald, J. & Smith, L. (2002). Applying Lee–Carter under conditions of variable mortality decline. *Population Studies* **56**, 325–336.
Brace, A., Gatarek, D. & Musiela, M. (1997). The market model of interest-rate dynamics. *Mathematical Finance* **7**, 127–155.
Brigo, D. & Mercurio, F. (2001). *Interest rate models: theory and practice*. Berlin: Springer.
Brouhns, N., Denuit, M. & Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected life tables. *Insurance: Mathematics and Economics* **31**, 373–393.
Cairns, A. J. G. (2004). *Interest rate models: an introduction*. Princeton, NJ: Princeton University Press.
Cairns, A. J. G. (2007). A multifactor generalisation of the Olivier–Smith model for stochastic mortality. *Proceedings of the 1st IAA Life Colloquium,* Stockholm.
Cairns, A. J. G., Blake, D. & Dowd, K. (2006a). Pricing death: frameworks for the valuation and securitization of mortality risk. *ASTIN Bulletin* **36**, 79–120.
Cairns, A. J. G., Blake, D. & Dowd, K. (2006b). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance* **73**, 687–718.
Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G., Epstein, D. & Khallaf-Allah, M. (2008). Mortality density forecasts: an analysis of six stochastic mortality models. Working paper, Heriot-Watt University, and Pensions Institute Discussion Paper PI–0801.
Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. & Balevich, I. (2007). A quantitative comparison of stochastic mortality models using data from England & Wales and the United States. Working paper, Heriot-Watt University, and Pensions Institute Discussion Paper PI–0701.
Chen, H. & Cox, S. H. (2007). Modeling mortality with jumps: transitory effects and pricing implication to mortality securitization. Working paper, Georgia State University.
Continuous Mortality Investigation (CMI) (2006). Stochastic projection methodologies: further progress and P-Spline model features, example results and implications. Working paper 20.
Continuous Mortality Investigation (CMI) (2007). Stochastic projection methodologies: Lee–Carter model features, example results and implications. Working paper 25.
Coughlan, G., Epstein, D., Sinha, A. & Honig, P. (2007). *q-forwards: derivatives for transferring longevity and mortality risks*. London: JPMorgan Pension Advisory Group, July.
Cowley, A. & Cummins, J. D. (2005). Securitization of life insurance assets and liabilities. *Journal of Risk and Insurance* **72**, 193–226.
Cox, J., Ingersoll, J. & Ross, S. (1985). A theory of the term-structure of interest rates. *Econometrica* **53**, 385–408.
Cox, S. H. & Lin, Y. (2007). Natural hedging of life and annuity risks. *North American Actuarial Journal* **11**, 1–15.
Currie, I. D., Durban, M. & Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* **4**, 279–298.
Czado, C., Delwarde, A. & Denuit, M. (2005). Bayesian Poisson log-linear mortality projections. *Insurance: Mathematics and Economics* **36**, 260–284.
Dahl, M. (2004). Stochastic mortality in life insurance: market reserves and mortality-linked insurance contracts. *Insurance: Mathematics and Economics* **35**, 113–136.
Dahl, M., Melchior, M. & Møller, T. (2008). On systematic mortality risk and risk minimisation with survivor swaps. *Scandinavian Actuarial Journal* **108**, 114–146.
Dahl, M. & Møller, T. (2006). Valuation and hedging of life insurance risks with systematic mortality risk. *Insurance: Mathematics and Economics* **39**, 193–217.
De Jong, P. & Tickle, L. (2006). Extending the Lee–Carter model of mortality projection. *Mathematical Population Studies* **13**, 1–18.
Delwarde, A., Denuit, M. & Eilers, P. (2007). Smoothing the Lee–Carter and Poisson log-bilinear models for mortality forecasting: a penalised log-likelihood approach. *Statistical Modelling* **7**, 29–48.
Dowd, K., Blake, D. & Cairns, A. J. G. (2007). Facing up to uncertain life expectancy: the longevity fan charts. Pensions Institute Discussion Paper PI–0703.
Dowd, K., Blake, D., Cairns, A. J. G. & Dawson, P. (2006). Survivor swaps. *Journal of Risk and Insurance* **73**, 1–17.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., Epstein, D. & Khalaf-Allah, M. (2008a). Evaluating the goodness of fit of stochastic mortality models. Forthcoming. Pensions Institute Discussion Paper PI–0802.
Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G. D., Epstein, D. & Khalaf-Allah, M. (2008b). Backtesting stochastic mortality models: an ex-post evaluation of multi-period-ahead density forecasts. Forthcoming. Pensions Institute Discussion Paper PI–0803.
Hainaut, D. & Devolder, P. (2008). Mortality modelling with Lévy processes. *Insurance: Mathematics and Economics* **42**, 409–418.
Heath, D., Jarrow, R. & Morton, A. (1992). Bond pricing the term structure of interest rates: a new methodology for contingent claims valuation. *Econometrica* **60**, 77–105.
Hull, J. & White, A. (1990). Pricing interest-rate derivative securities. *Review of Financial Studies* **3**, 573–592.
Jacobsen, R., Keiding, N. & Lynge, E. (2002). Long-term mortality trends behind low life expectancy of Danish women. *Journal of Epidemiology Community Health* **56**, 205–208.
Jamshidian, F. (1997). LIBOR and swap market models and measures. *Finance and Stochastics* **1**, 293–330.
Karatzas, I. & Shreve, S. E. (1998). *Methods of mathematical finance*. New York: Springer.
Kirkby, J. G. & Currie, I. D. (2007). Smooth models of mortality with period shocks. *Proceedings of 22nd International Workshop on Statistical Modelling*, Barcelona, 374–379.
Lee, R. D. & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association* **87**, 659–675.
Lee, R. & Miller, T. (2001). Evaluating the performance of the Lee–Carter model for forecasting mortality. *Demography* **38**, 537–549.
Lin, Y. & Cox, S. H. (2005). Securitization of mortality risks in life annuities. *Journal of Risk and Insurance* **72**, 227–252.
Loeys, J., Panigirtzoglou, N. & Ribeiro, R. M. (2007). Longevity: a market in the making. Technical report, J. P. Morgan Securities Ltd.
Milevsky, M. A. & Promislow, S. D. (2001). Mortality derivatives and the option to annuitise. *Insurance: Mathematics and Economics* **29**, 299–318.
Miltersen, K. R. & Persson, S.-A. (2005). Is mortality dead? Stochastic forward force of mortality determined by no arbitrage. Working paper, University of Bergen.
Miltersen, K. R., Sandmann, K. & Sondermann, D. (1997). Closed-form solutions for term structure derivatives with log-normal interest rates. *Journal of Finance* **52**, 409–430.
O'Brien, C. (1999). Actuaries and accountability. Presidential Address to the Manchester Actuarial Society, 1999.
Olivier, P. & Jeffery, T. (2004). Stochastic mortality models. Presentation to the Society of Actuaries of Ireland. Available at: www.actuaries.ie (Events and Papers) (accessed 9 July 2008).
Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology* **14**, 124–129.
Renshaw, A. E. & Haberman, S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics* **33**, 255–272.
Renshaw, A. E. & Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* **38**, 556–570.
Richards, S. J., Kirkby, J. G. & Currie, I. D. (2006). The importance of year of birth in two-dimensional mortality data, (with discussion). *British Actuarial Journal* **12**, 5–61.
Schrager, D. F. (2006). Affine stochastic mortality. *Insurance: Mathematics and Economics* **38**, 81–97.
Smith, A. D. (2005). Stochastic mortality modelling. Workshop on the Interface between Quantitative Finance and Insurance, International Centre for the Mathematical Sciences, Edinburgh. Available at: www.icms.org.uk/archive/meetings/2005/quantfinance/ (accessed 9 July 2008).
Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**, 177–188.
Willets, R. C. (2004). The cohort effect: insights and explanations. *British Actuarial Journal* **10**, 833–877.

Appendix: Identifiability constraints

For the model discussed in Section 4.4 there is an identifiability problem. This model, as with others such as the Lee–Carter model requires the use of identifiability constraints to ensure that the estimation problem is well specified. For this model we have the following identifiability problem. Suppose that the best-fitting set of parameters are $\hat{\kappa}_t^{(1)}$, $\hat{\kappa}_t^{(2)}$, $\hat{\kappa}_t^{(3)}$, and $\hat{\gamma}_{t-x}^{(4)}$. Now consider the alternative parameterization for constants ϕ_1, ϕ_2 and ϕ_3:

$$
\begin{aligned}
&\tilde{\gamma}^{(4)}_{t-x} = \hat{\gamma}^{(4)}_c + \phi_1 + \phi_2(t-x) + \phi_3(t-x)^2 \\
&\tilde{\kappa}^{(1)}_t = \hat{\kappa}^{(1)}_t + \psi_1(t) \\
&\tilde{\kappa}^{(2)}_t = \hat{\kappa}^{(2)}_t + \psi_2(t) \\
&\tilde{\kappa}^{(3)}_t = \hat{\kappa}^{(3)}_t + \psi_3(t) \\
\text{where}\quad &\psi_1(t) = -\phi_1 - \phi_2(t-\bar{x}) - \phi_3(t-\bar{x})^2 - \phi_3\sigma_x^2 \\
&\psi_2(t) = \phi_2 + 2\phi_3(t-\bar{x}) \\
&\psi_3(t) = -\phi_3.
\end{aligned}
$$

With a bit of tedious algebra, it can be shown that the two parameterizations give the same values for $q(t, x)$ for all t and x.

To avoid this problem, we need to introduce three constraints that make the arbitrary adjustments above impossible. A simple approach would be to require that $\Sigma_t \kappa^{(i)}_t = 0$ for $i = 1, 2, 3$. However, for reasons given below we propose an alternative set of constraints:

$$
\sum_{c=c_0}^{c_1} \gamma^{(4)}_c = 0
$$

$$
\sum_{c=c_0}^{c_1} c\gamma^{(4)}_c = 0
$$

$$
\sum_{c=c_0}^{c_1} c^2\gamma^{(4)}_c = 0
$$

where c_0 and c_1 are the earliest and latest years of birth to which a cohort effect is fitted. The reason for this choice is that if we use least squares to fit a quadratic function $\phi_1 + \phi_2 c + \phi_3 c^2$ to $\gamma^{(4)}_c$, the constraint ensures that $\hat{\phi}_1 = \hat{\phi}_2 = \hat{\phi}_3 = 0$. This means that the fitted $\gamma^{(4)}_c$ process will fluctuate around zero and it will have no discernible linear trend or curvature (as can be seen in Figure 6).