THE IMPORTANCE OF YEAR OF BIRTH IN TWO-DIMENSIONAL MORTALITY DATA

By S. J. Richards, J. G. Kirkby and I. D. Currie

[Presented to the Institute of Actuaries, 24 October 2005]

ABSTRACT

Late life mortality patterns are of crucial interest to actuaries assessing longevity risk. One important explanatory variable is year of birth. We present the results of various analyses demonstrating this, including a statistical model which lends weight to the importance of year-of-birth effects in both population and insured data. We further find that a model based on age and year of birth fits United Kingdom mortality data better than a model based on age and period, suggesting that cohort effects are more significant than period effects. The financial implications of these cohort effects are considerable for portfolios with long-term longevity exposure, such as annuities written by insurance companies and defined benefit pension schemes.

KEYWORDS

Mortality Improvements; Cohorts; Early Life Conditions; Moving Averages; Kernel Smoothing; Penalised Splines

CONTACT ADDRESS

Stephen Richards, 4 Caledonian Place, Edinburgh EH11 4AS, U.K.
Web: www.richardsconsulting.co.uk; Tel: +44(0)131 315 4470;
E-mail: stephen@richardsconsulting.co.uk

1. Introduction

1.1 A macro-economic environment of low inflation and modest nominal investment returns has made increasing longevity a much bigger issue for sponsors of defined benefit pension schemes. Low nominal interest rates have also thrown the spotlight onto the mortality assumptions used by insurers for fixed annuity business. Furthermore, new product innovations, such as equity release mortgages, have created entirely new forms of financial exposure to longevity risk (Richards & Jones, 2004). The need to understand late-life mortality patterns, especially future mortality improvements, has grown rapidly in recent years.

1.2 As shown by Willets (1999), an important explanatory variable for mortality patterns is year of birth, and a key task is the efficient separation of

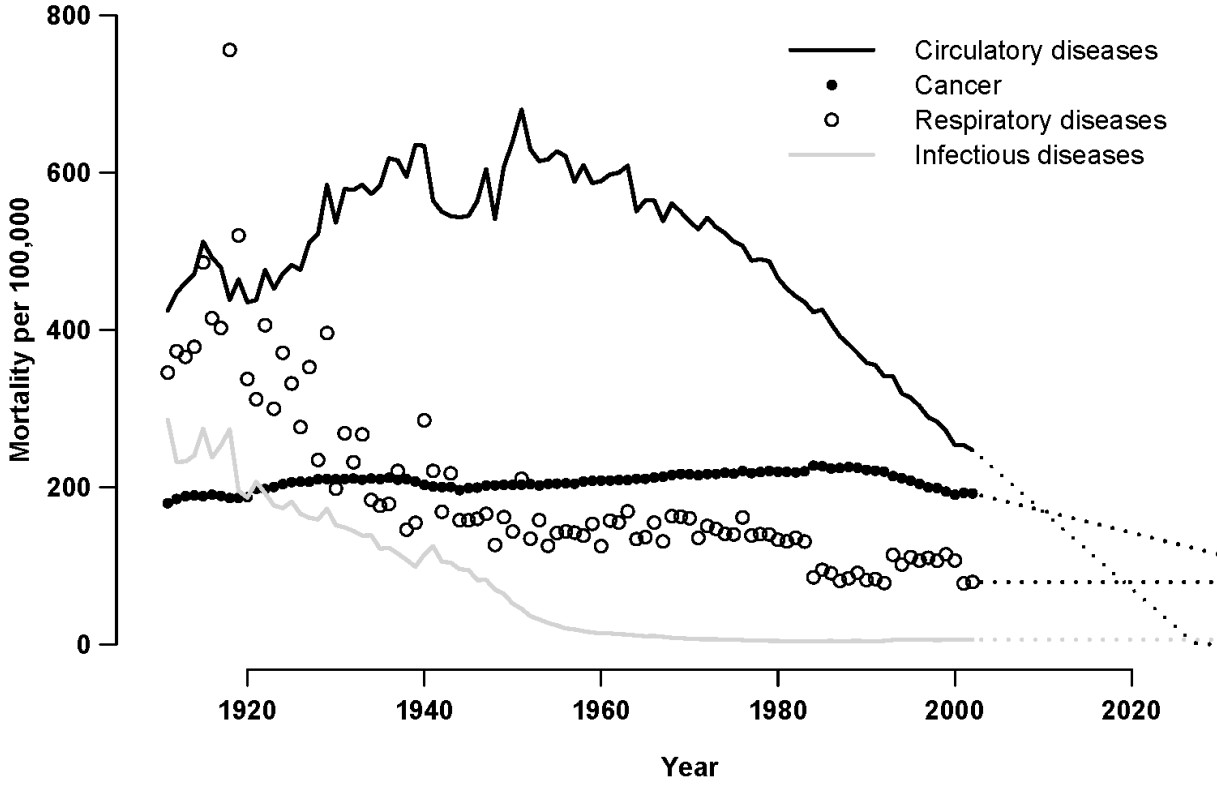
patterns from background noise. This paper presents the results of three different techniques for achieving this with two-dimensional mortality data. In each case, the year of birth proves to be a strong explanatory factor for patterns of mortality improvement in the United Kingdom, both in the insured and in the general populations. A statistical method is presented, which formally shows the dominance of cohort effects over period effects in the U.K.

1.3 It is not the goal of this paper to comment on work done by the Continuous Mortality Investigation Bureau (CMIB), although the interested reader is directed to CMIB (2005), which contains further consideration of penalised-spline models for projection purposes.

2. Factors Driving Mortality Patterns by Year of Birth

2.1 Since the 1970s, new evidence has been gathered which suggests that adult chronic disease is at least partly rooted in conditions experienced in early life, including conditions *in utero*. Ongoing research uses a variety of means of measuring this. One simple approach uses the month or season of birth as a proxy for the pre-natal and perinatal environmental conditions experienced by a child — see Gavrilov & Gavrilova (1999), Doblhammer & Vaupel (2001) and Gavrilova *et al.* (2002). Other work finds that disease load in the first year of life has a strong impact on mortality in later life, especially diseases from airborne infections (Bengtsson & Lindstrom, 2000). The theory is that infection during crucial developmental stages is damaging, and not just in direct mortality terms. For example, much research has been devoted to the alleged connection between influenza infection and schizophrenia: see Izumoto *et al.* (1999) and Selten *et al.* (1999). Influenza is increasingly suspected of playing a major role in susceptibility to heart disease (Madjid *et al.*, 2004), which remains the leading cause of death amongst the elderly in the U.K. Some mechanisms and results are discussed by Gluckman & Hanson (2004), while Finch & Crimmins (2004) discuss the idea of a 'cohort morbidity phenotype'. Under this hypothesis, early-life exposure to infectious agents causes an inflammatory reaction which can lead to chronic disease in middle age. Figure 1 shows the substantial fall in mortality due to infectious agents and respiratory disease in the first half of the twentieth century in England and Wales. Perhaps not coincidentally, mortality from circulatory disease shows a steep and sustained fall as the generation born in the first half of the twentieth century enters middle age in the latter half.

2.2 For ethical reasons, it is impossible to conduct experiments to conclusively verify the causal pathways in humans. However, certain historical events do enable some hypotheses to be researched and tested. One example is the famine in the northern Netherlands in 1944 - 45, during the



Source: ONS data with own extrapolations beyond 2002

Figure 1. Mortality rates per 100,000 for England and Wales by main causes of death

closing stages of the Second World War. This gave, not only an exposed population which is precisely known, but also a comparable contemporaneous control population in the rest of the Netherlands, which did not suffer the famine. Analysis of the mortality of those born before, during and after the famine gives insight into the relationship between under nutrition and late-life susceptibility to disease (Rosebloom *et al.*, 2001). The results are not conclusive, however, while the Dutch famine of 1944 - 45 appeared to lead to increased susceptibility to coronary heart disease (CHD), the famine associated with the Leningrad Siege of 1941 did not. Stanner *et al.* (1997) found that the Leningrad group had greater levels of obesity and higher blood pressure for those exposed to intrauterine malnutrition than those either unexposed to malnutrition, or else exposed to malnutrition only as infants. This suggests that it is not merely early life conditions which drive later health and mortality, but also the conditions which applied at key developmental stages.

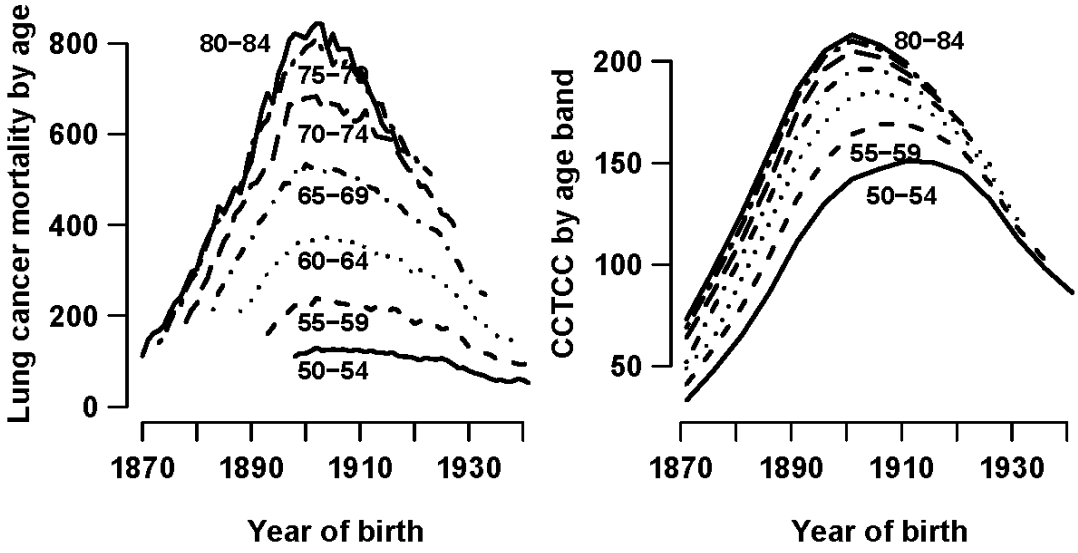
2.3 Early-life conditions impact directly, not only on later health and mortality, but also act as contributing factors towards wealth accumulation and socio-economic status. Links between mortality and wealth, or between mortality and socio-economic status, cannot be interpreted in isolation. Some

research does a control for socio-economic variation (McCarron *et al.*, 2002), and yet still finds a strong inverse relationship between height (an indicator of early life conditions) and increased cardiovascular mortality. A possible mechanism for this is that low birth weight is associated with concentric enlargement of the left ventricle in adult life (Vijayakumar *et al.*, 1995).

2.4 Some studies compare the relative influence of year of birth and year of observation. Age-period-cohort studies in Sweden suggest that both contain relevant information explaining trends in stroke mortality (Peltonen & Asplund, 1996). In some instances, year of birth proves to be more important than year of observation. For example, stillbirth rates have been linked to sub-optimal early life conditions, and there is evidence in Norway that stillbirth rates are more closely linked to the mother's own year of birth than the year of the stillbirth itself (Liestol, 1981). One interpretation here might be that the mother's early-life conditions dominate the effect of improved standards of living and medical care. This makes sense when one notes that human females are born with all the oocytes (eggs) they will ever have, making pre-natal conditions decisive, both for a woman's future fertility and her mortality. One could even imagine using cohort stillbirth data as a signpost for late-life cohort mortality patterns in females.

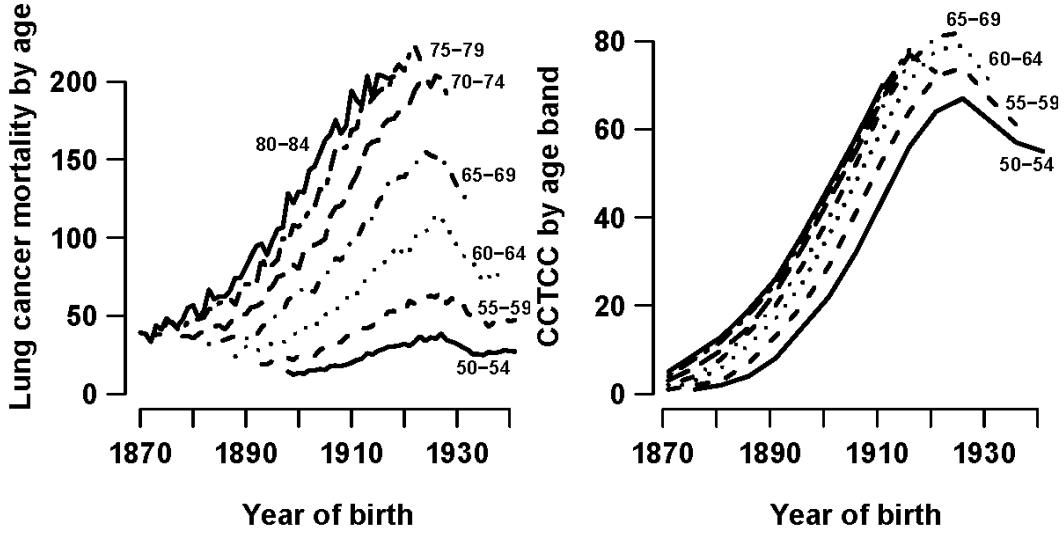
2.5 Besides pre-natal and early-life conditions, mortality is also strongly influenced by lifetime health behaviours. A great deal is now known about mechanisms linking various health behaviours to mortality outcomes, especially smoking (Doll & Hill, 1954; Doll *et al.*, 2004) and diet (Ames, 1998). Where there are pronounced and consistent generational differences in such life course health behaviours, these will also be picked up in a year-of-birth analysis. An example of this is smoking and rates of lung cancer mortality, as shown in Figures 2(a) and 2(b). When one looks at these graphs, and also Figure 1, they suggest that the cohort effect in the U.K. is perhaps less about a 'healthy generation benefiting from wartime rationing and the Welfare State', and rather more the result of preceding generations being particularly unhealthy and, indeed, 'damaged'.

2.6 It has been suggested that the cohort effect is largely due to the changing incidence in smoking. It is true that changes in personal smoking habits are a component part of the cohort effect, as evidenced in Figures 2(a) and 2(b), and confirmed by Willets (2004). However, smoking does not account for the cohort effect in full; Willets (2004) found that only part of the cohort effect could be explained by lung cancer incidence (and thus smoking), whereas Doll *et al.* (2004) found 'a progressive reduction in the mortality of never-smokers', with the survival probability of reaching age 90 from age 70 increasing threefold amongst those who had never smoked over the 50-year study period. Doll *et al.* (2004) found year-of-birth patterns in mortality rates amongst lifelong non-smokers, as shown in Figure 2(c), suggesting that smoking patterns can be only a partial explanation of the cohort effect.



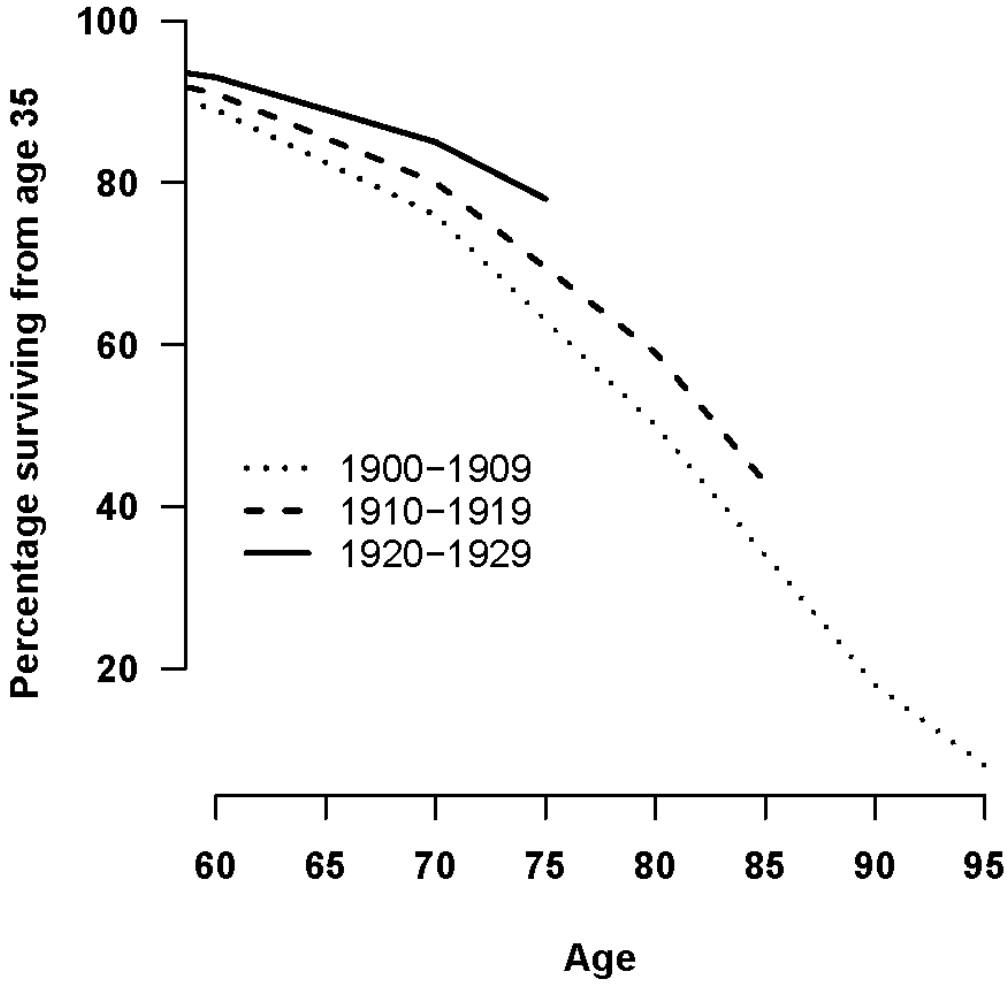
Source: Lee *et al.* (1990), Forey *et al.* (1993) and ONS (Twentieth Century Mortality)

Figure 2(a). Lung cancer mortality rates per 100,000 for U.K. males by age band (left) and estimated cumulative lifetime cigarette consumption for males (right); CCTCC = Cumulative Constant Tar Cigarette Consumption, a standardised measure of lifetime cigarette consumption, shown in units of 100,000



Source: Lee *et al.* (1990), Forey *et al.* (1993) and ONS (Twentieth Century Mortality)

Figure 2(b). Lung cancer mortality rates per 100,000 for U.K. females by age band (left) and estimated cumulative lifetime cigarette consumption for females (right); age band labelling is as per Figure 2(a) for males; CCTCC = Cumulative Constant Tar Cigarette Consumption, a standardised measure of lifetime cigarette consumption, shown in units of 100,000



Source: Doll *et al.* (2004)

Figure 2(c). Cohort survival rates for lifelong non-smokers among British male doctors; clear differences exist between the survival rates of lifelong non-smokers, suggesting that the cohort effect is driven by more than just personal smoking habits

2.7 The subject of initial damage done to an individual's DNA is also a key topic. Gavrilov & Gavrilova (2001) have developed a simple, yet elegant, theory of human ageing in engineering terms: the progressive failure of redundancy as the driver for the shape of the curve of human mortality rates. Gavrilov & Gavrilova (2004) further show the theoretical effects of high initial damage load (HIDL) and the idea of an initial virtual age as a means of explaining how some individuals are born further into the ageing process than others. This idea is immediately familiar to actuaries as the long-established practice of rating lives by treating them as being younger or older than their chronological age for the purposes of pricing and reserving. Gavrilov & Gavrilova not only developed the reliability theory of ageing, but also identified precise stages at which the initial damage can occur: from

conception (Gavrilov *et al.*, 1997; Gavrilov & Gavrilova, 2000); through to inter-uterine conditions (Gavrilova *et al.*, 2002; Gavrilov & Gavrilova, 2003); and even the process of birth itself, which causes hypoxia and consequent DNA damage.

2.8 Year of birth is therefore a potentially powerful combined proxy for fœtal and gestational conditions, early-life disease load and lifetime patterns of health behaviours. This paper is concerned with measuring the extent to which year-of-birth effects explain observed mortality improvement patterns in U.K. males. Willets (2004) gives further references and discussion of year-of-birth effects in the U.K. We note that there are other, more significant rating factors for mortality — most obviously socio-economic group — but we do not have the data to consider their impact. The relative importance of various rating factors in annuities and pensions business is explored in more detail in Richards & Jones (2004).

3. Estimating Year-of-Birth Effects using Moving Averages

3.1 The Government Actuary's Department (GAD) calculates mortality rates and life tables based on data on numbers of deaths and population estimates produced by the Registrars General of England and Wales, Scotland and Northern Ireland. The GAD has two data sets. The first is in the form of interim life tables, which cover individual ages from zero to 100, and separately cover the four territories of the U.K. (England, Wales, Scotland and Northern Ireland). An interim life table is based on the experience of a triennium, the most recent available being the calendar years 2001 - 2003. Interim life tables are publicly available from the GAD website for triennia from 1980 onwards.

3.2 The second GAD data set differs from the first in that it is not publicly available, only covers England and Wales combined, and only covers individual ages from 20 to 89. However, this second data set has two important advantages: it is annual, not triennial, data; and the data stretch back to 1961, rather than to 1980.

3.3 Both sets of data give the central exposure to risk, so we are handling central rates of mortality (m_x), not probabilities (q_x). For the purposes of this paper, we will use both data sets: the triennial data for illustrative, one-dimensional work; and the annual data for the main, two-dimensional work. The data are made available by attained age and year of observation, but we will present our results by year of birth and year of observation to compare and contrast the two. Figure 3 shows the transformation of the data that is necessary for cohort-orientated work, and the resultant incomplete information which results for the earlier and later cohorts.

3.4 As championed by Willets (1999) and used in CMIB (2002), rather

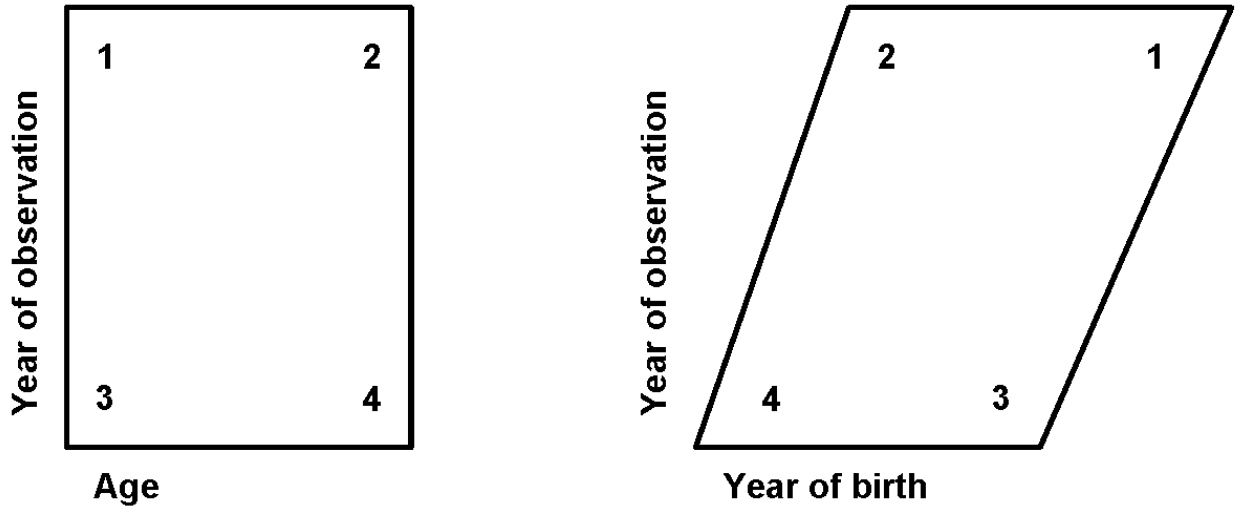


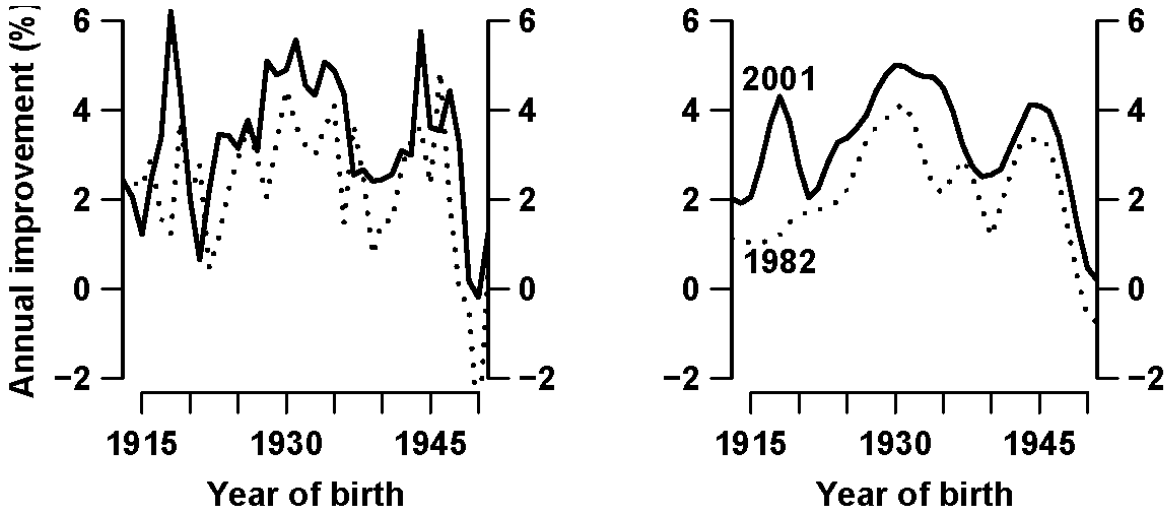
Figure 3. Transformation of data layout from year of observation against age (left) to year of observation against birth cohort (right); the four numbers show how a given mortality rate for age and year of observation relocates when restructuring the data on a cohort basis

than use the central mortality rates $m_{x,t}$, we will work with the mortality improvements $\Delta m_{x,t}$, where:

$$\Delta m_{x,t} = 1 - \frac{m_{x,t}}{m_{x,t-1}}$$

where x is the attained age and t is calendar time. To see underlying patterns, it is necessary to remove the effects of random variation. This we can achieve by smoothing, and Figure 4 shows how one-dimensional smoothing uses information at adjacent ages to both damp down random variation and to 'borrow' adjacent information in determining underlying patterns. The left panel in Figure 4 shows the raw, unsmoothed improvements, while the right panel shows the resulting smoothed values after using a simple moving average. The effectiveness of the smoothing function can be seen quite clearly, as it picks out consistent year-of-birth patterns for two quite different triennia, separated by nearly two decades. One must bear in mind that comparability is between lines, not necessarily within the lines themselves; the year-of-birth figures for 1930 are for an age ten years older than the figures for 1940.

3.5 The right panel in Figure 4 contains several useful pieces of information. Firstly, we can clearly see the much documented 'cohort effect', whereby very strong mortality improvements are exhibited by the generations born around 1931 (Willets, 2004). Secondly, mortality improvements appear to increase with advancing age; the line of improvements during 2001 is uniformly higher than the line a decade earlier



Source: Own calculations using GAD data

Figure 4. Unsmoothed (left) and smoothed (right) mortality improvements for males in the U.K.; triennium-to-triennium improvements for 1980 - 82 to 1981 - 83 and 1999 - 2001 to 2000 - 2002 (labelled 1982 and 2001 after the mid-point of the later triennium in each case); mortality improvements are smoothed using a simple (1, 2, 3, 2, 1) moving average centred on each year of birth, divided by nine to normalise the smoothed values; this simple moving average function was chosen for its fit by eye, rather than by a fitting procedure; similar patterns exist for females (not shown), as discussed in Richards & Jones (2004)

for all but the extremely elderly (although this could simply be a period effect, say, due to the milder winters of 2000 - 2002). Thirdly, a second cohort with strong mortality improvements is also clearly visible for those males born just before 1945.

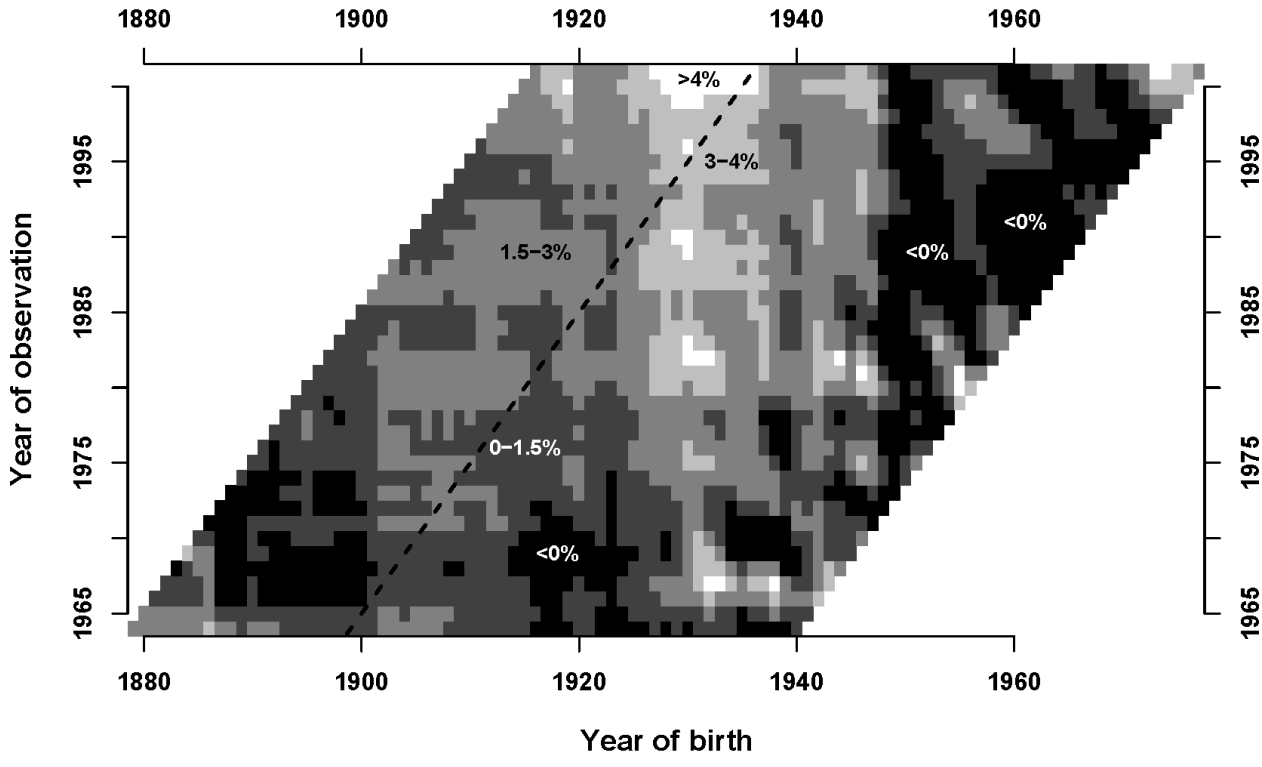
3.6 Despite the caveat about not comparing intra-line patterns, there is, nevertheless, something about those born around 1939 that gives them materially lower mortality improvements than the generations both before and after them. Note, however, the distinction between low relative mortality improvements and worsening mortality; it is only when mortality improvements are negative that mortality is actually worsening from generation to generation. The generation of 1939 may have lower relative improvements than the generation of 1931, but the former will still live longer. A generation 'banks' all the improvements of the generations which precede it. A cohort with low-but-positive improvements still has lower mortality rates than the cohort which precedes it.

3.7 A broadly similar pattern exists for females, but with two notable exceptions (not shown). Firstly, the gap between the triennial improvements is much smaller. Secondly, the dip for the cohort born around 1939 is much

less pronounced, so much so that it appears non-existent for the 2001 curve (the 2000 - 2002 triennium). Gavrilova *et al.* (2002) discuss the biological basis for females being less susceptible than males to adverse conditions. If variations in adverse life-course conditions drive cohort effects even partly, then the strength of cohort effects will be greater for males.

3.8 Smoothing works by damping down random variation and borrowing information from adjacent points. Working simultaneously in two dimensions will, therefore, give greater insight into patterns of mortality improvement; instead of two adjacent points from which to borrow information, there are eight. Figure 5 shows the result of such two-dimensional smoothing.

3.9 The highest improvement rates are white and the lowest improvement rates are black, with various scales of grey in between. If you think in terms of mountains, the snow covered peaks are the high improvement rates and the dark valleys are the low improvement rates. The block-like pattern is an artefact of the moving average smoothing. The white triangles at the top left and bottom right are artefacts of the data set: data are not consistently



Source: Own calculations using GAD data

Figure 5. Smoothed male mortality improvements in England and Wales, displayed using five grey scale levels; year-on-year improvements for individual years; the dashed line connects improvements at age 65, which has been the U.K. state pension age for males throughout; note the missing triangles of data caused by the transformation in Figure 3

available above age 84 (causing missing data in top left); nor were they available for this work under age 20 (causing missing data in the bottom right). The weights for the two-dimensional moving average used in Figure 5 are given below. In order to avoid edge effects, smoothed values are not attempted where any necessary surrounding data point is missing:

$$\frac{1}{81}\begin{pmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 4 & 6 & 4 & 2 \\ 3 & 6 & 9 & 6 & 3 \\ 2 & 4 & 6 & 4 & 2 \\ 1 & 2 & 3 & 2 & 1 \end{pmatrix}.$$

3.10 Interestingly, it is a smaller number of grey scale levels which best brings out the broad sweep of underlying pattern; year-of-birth effects (vertical patterns) appear more dominant than year-of-observation effects (horizontal patterns). The low and negative improvements around year of birth 1950 are both stark and persistent across the 20-year period covered. Using a larger number of grey scale levels allows the identification of much narrower year-of-birth patterns, however. This topic of 'tuning' the smoothing will be returned to in the following sections on kernel and penalised spline smoothing.

3.11 One other feature suggested by Figure 5 was first hinted at in Figure 4; the acceleration of mortality improvements with advancing age. This is best illustrated for birth years up to around 1940; there is an apparent tendency for mortality improvements to be higher (lighter colours) as calendar time and age increase.

3.12 Revealing though these graphs are, there are some fundamental limitations of the moving-average approach which must be noted. Firstly, although the moving average has revealed the possible existence of strong year-of-birth (cohort) effects, it does not provide a framework for testing this statistically.

3.13 Secondly, the smoothing function used applies only to points which are co-incident with the data. This yields the block-like effect in the two-dimensional graph. To get intermediate smoothed values, we would require a separate set of smoothing functions, and thus different weights to those in ¶3.9. A regression method is needed for continuous smoothed values.

3.14 Thirdly, although the moving-average approach could be extended to provide future projections, it is clumsy. It also cannot provide a cohesive statistical framework with confidence intervals for those projections. While a projection method could be conceived of, it would not be integrated with the smoothing function, i.e. the projection parameters would not be set with reference to any kind of regression model underlying the smoothing.

3.15 Finally, the weights used here were set 'by eye', but could have

been set less arbitrarily, e.g. with some variant of the 'least squares' approach. However, this would not have a statistical framework, and therefore would not have yielded any formal statistical badness-of-fit test for the smoothing.

4. Estimating Year-of-Birth Effects using Kernel Smoothing

4.1 Although moving averages are fine for initial exploration, we can improve our work by using kernel smoothing. Here, smoothed values for $\Delta m_{x,y,t}$ are calculated from a weighted average of the data values $\Delta m_{x_i,y_i,t_i}$, where i indexes the available data points, x is the attained age, y is the birth cohort and t is calendar time. The relative mortality improvement is defined as $\Delta m_{x,y,t} = 1 - \dfrac{m_{x,y,t}}{m_{x,y-1,t-1}}$. The weights for the averaging process are given by

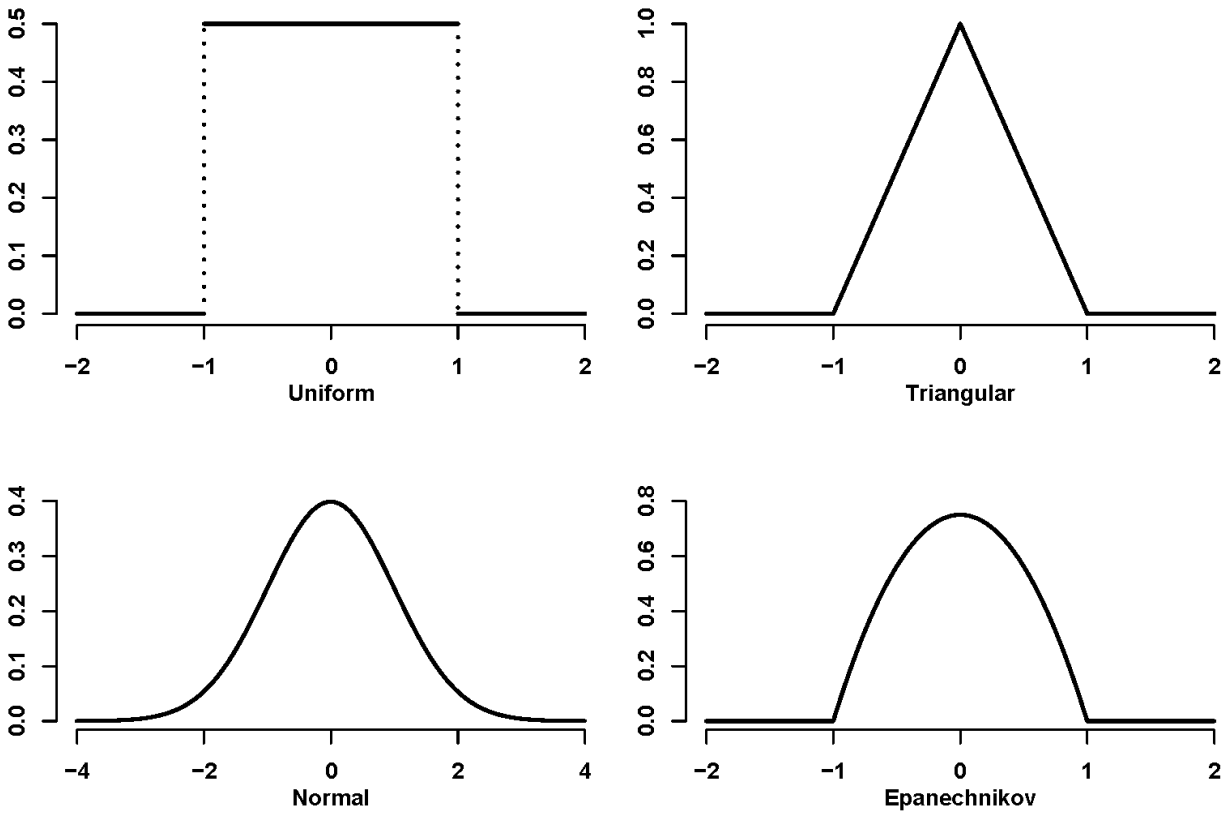
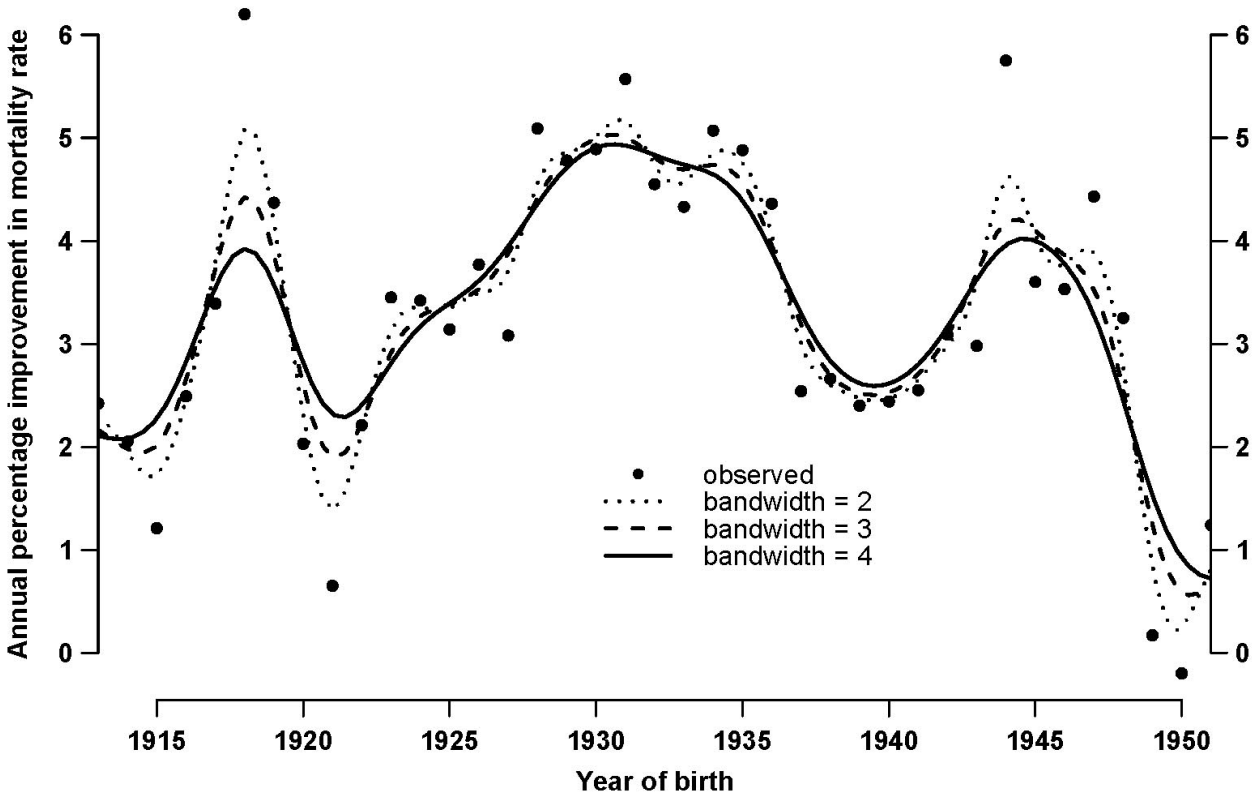


Figure 6. Some kernel functions; kernel efficiency is measured relative to the Epanechnikov kernel, the most efficient with $f(x) = \frac{3}{4}(1 - x^2)$, $-1 \leq x \leq 1$, otherwise zero; the Normal (or Gaussian) kernel, for example, has an efficiency within five per cent of the maximum; even the uniform kernel is only seven per cent lower than maximum, confirming the dominant role of the bandwidth parameter in kernel smoothing



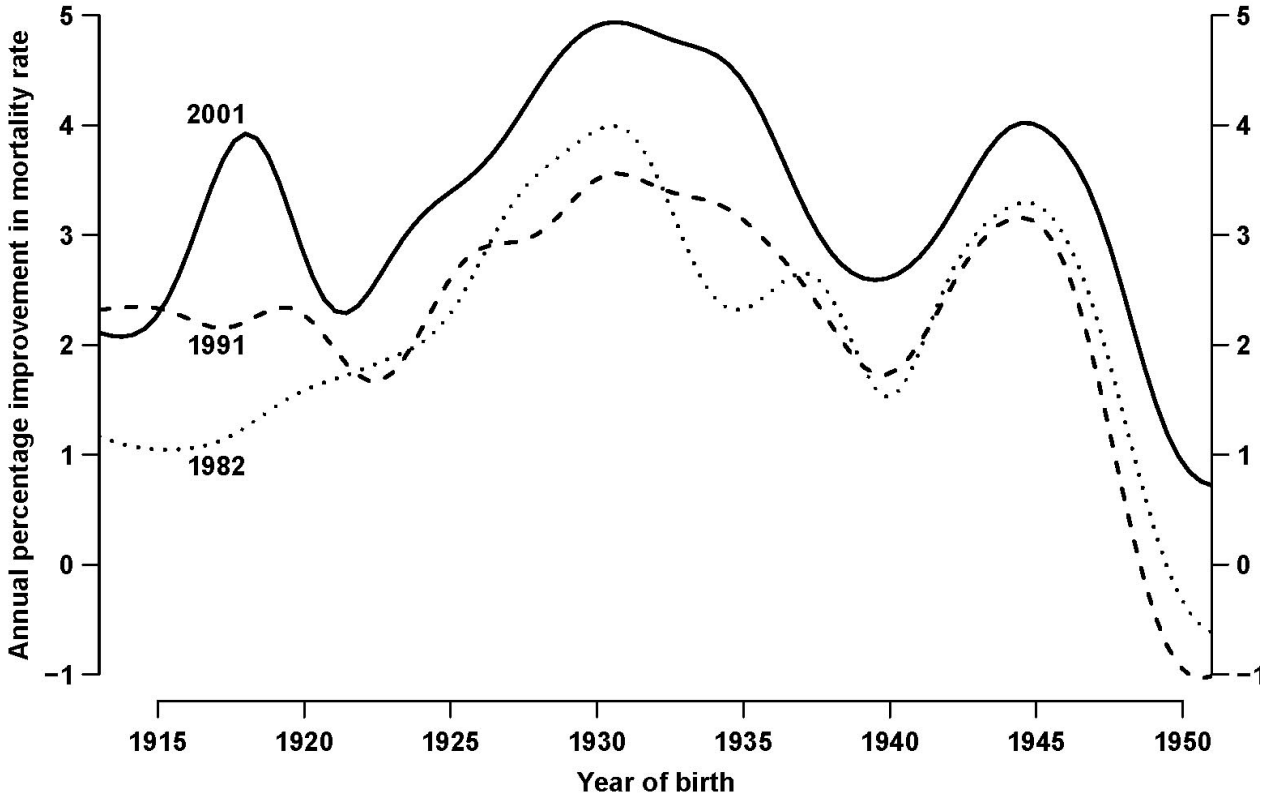
Source: Own calculations using GAD data

Figure 7. Male mortality improvement rates by year of birth in the U.K. from triennium 1999 - 2001 to triennium 2000 - 2002 with kernel-smoothed trend lines (Nadaraya-Watson kernel smoother with Gaussian kernel function and the bandwidth parameters shown)

the kernel function f, which returns a weight based on the distance of the data point from the fitting point. Kernel smoothing takes place over an arbitrarily fine grid. Together with a function delivering continuously varying weights based on distance, kernel smoothing can eliminate the distracting block-like patterns of Figure 5.

4.2 Some sample one-dimensional kernel functions are given in Figure 6. Kernel functions are symmetric about zero and integrate to unity. It is usual for the function maximum to occur at zero, with values decaying monotonically for values progressively further from zero (one exception is the little-used uniform kernel in Figure 6). Kernel functions vary in their efficiency, which is measured as the asymptotic mean integrated squared error (AMISE). Kernel efficiency is expressed relative to the AMISE of the Epanechnikov kernel, since this kernel minimises the AMISE, and is therefore optimal. More details on kernels, their efficiency and their computation, can be found in Wand & Jones (1995).

4.3 The kernel function supplies weights to apply to nearby data points, based on their distance from the central smoothing point. This distance is



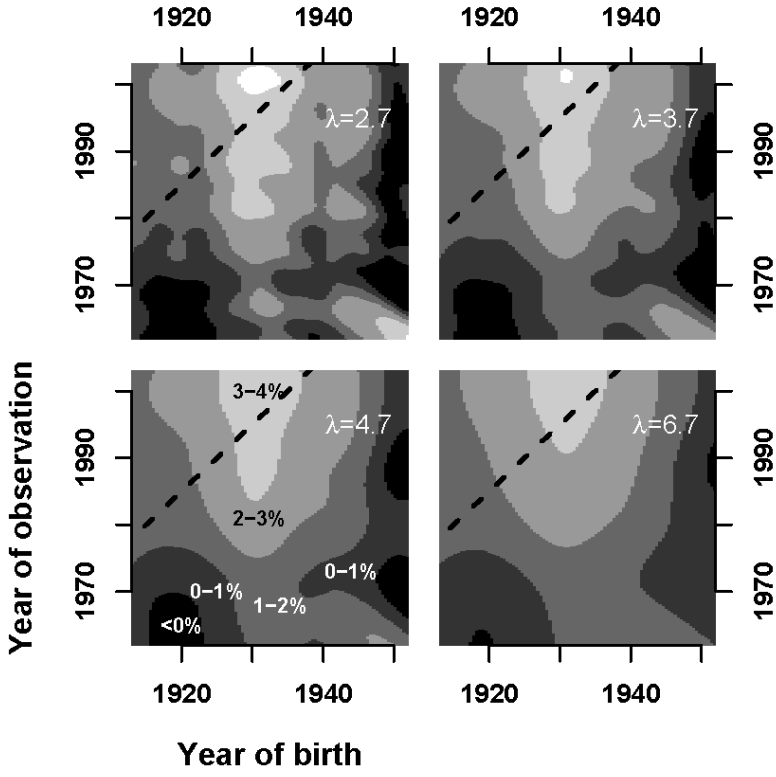
Source: Own calculations using GAD data

Figure 8. Kernel smoothed male mortality improvement rates by year of birth in the U.K. for three triennia centred on 1982, 1991 and 2001 (Nadaraya-Watson kernel smoother with Gaussian kernel function and with a bandwidth parameter of four)

adjusted by the bandwidth parameter λ, which can be thought of as the degree of smoothing strength; the higher the value of λ, the stronger the smoothing being applied. The choice of kernel function is less important, as it is the bandwidth parameter λ which makes the greatest difference, as demonstrated later in Figure 7, and again later in Figure 9.

4.4 By way of illustration, the smoothing in Figures 7 and 8 is one-dimensional across year of birth. Each smoothed value $\Delta m^s_{x,y,t}$ is a weighted combination of the data points $\Delta m_{x_i,y_i,t_i}$. The weight for each $\Delta m_{x_i,y_i,t_i}$ is $f\left(\frac{y-y_i}{\lambda}\right)$.

4.5 A kernel function can be extended to two dimensions by rotating the shapes in Figure 6 around the axis $x = 0$. The two-dimensional kernel smoothed weights for calculating $\Delta m^s_{x,y,t}$ in Figures 9, 10 and 11, are calculated using weights from $f\left(\frac{y-y_i}{\lambda}, \frac{t-t_i}{\lambda}\right)$ applied to $\Delta m_{x_i,y_i,t_i}$. Further details on kernel smoothing can be found in Wand & Jones (1995).



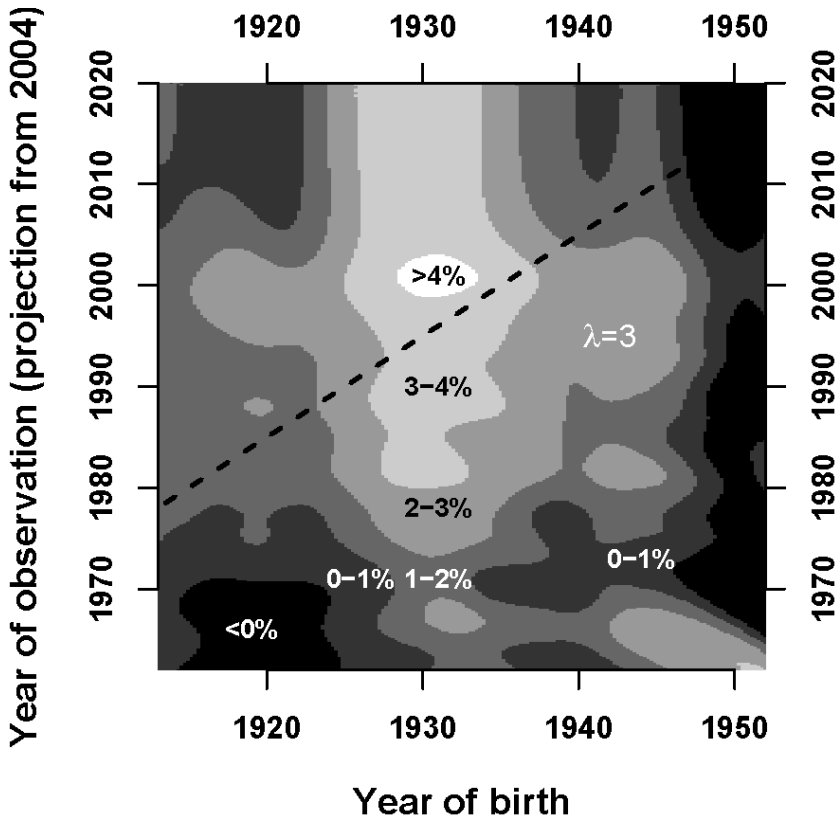
Source: Own calculations using GAD data

Figure 9. Smoothed male mortality improvements in England and Wales; year-on-year improvements for individual years, smoothing using Nadaraya-Watson two-dimensional kernel smoother and Gaussian kernel function with given bandwidth parameter λ; the dashed line connects age 65 in each case; common colour break levels are set as marked in the lower left graph; the pure white areas in the top two graphs represent improvements in excess of 4% p.a.

4.6 As it happens, a moving average is simply a special case of kernel smoothing where the smoothing grid exactly coincides with the data, i.e. with no intermediate points. The one-dimensional moving average used for Figure 4 is essentially a discrete triangular kernel, while the two-dimensional moving average used for Figure 5 is essentially a discrete pyramidal kernel.

4.7 We have used the Nadaraya-Watson kernel smoother (Nadaraya, 1964; Watson, 1964). In addition to the choice of kernel and bandwidth parameter, we must also select a grid over which the kernel smoothing takes place. It is desirable to choose a grid with more points than data in order to avoid the distracting blocks which arise with moving averages, as in Figure 5. For the two-dimensional smoothing in this paper, we have used a square grid of 512 points on each side.

4.8 The full power of the bandwidth parameter becomes evident in the two-dimensional case in Figure 9. A narrow bandwidth picks up very specific year-of-birth effects in the top left panel. Progressively wider bandwidths



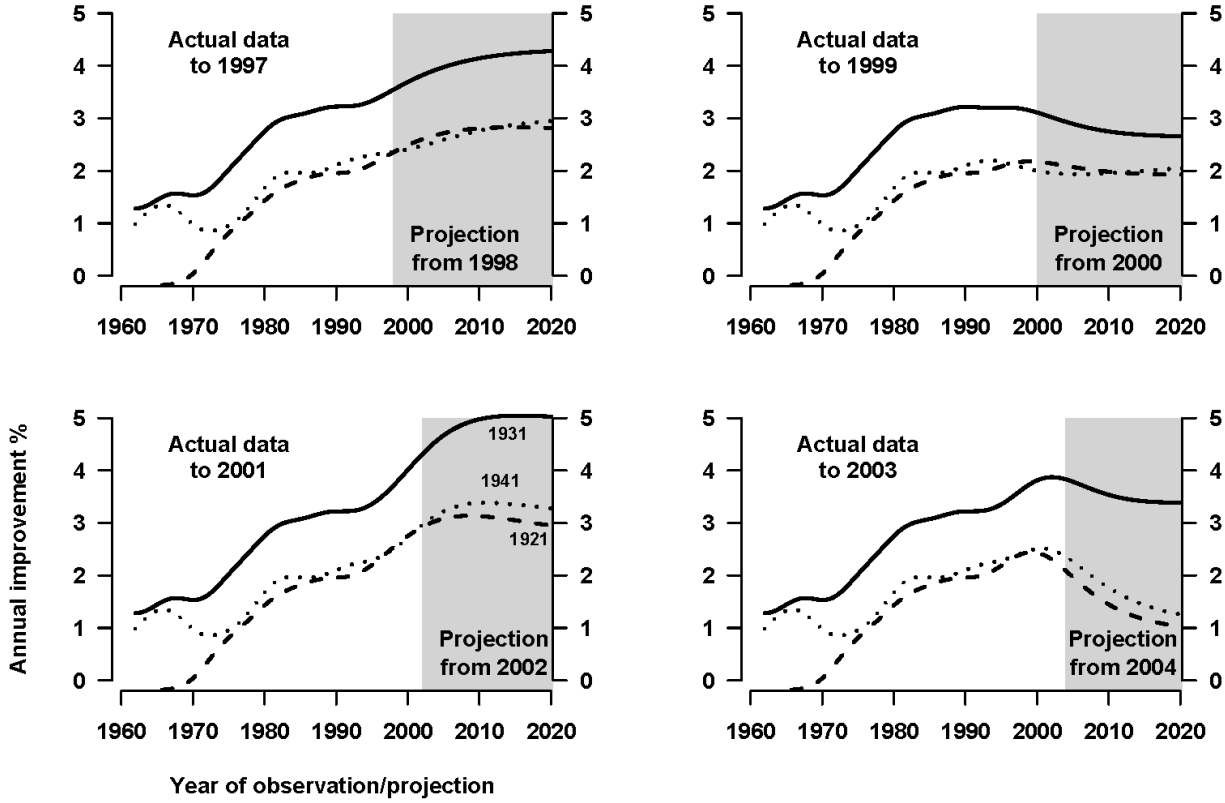
Source: Own calculations using GAD data

Figure 10. Smoothed male mortality improvements in England and Wales with projection; year-on-year improvements for individual years, smoothed using the Nadaraya-Watson two-dimensional kernel smoother with the Gaussian kernel function and a bandwidth parameter $\lambda = 3.0$; the dashed line represents age 65

fail to pick very narrow year-of-birth effects, but become much better able at identifying the broad sweep of trends. Note that the kernel smoother knows nothing of age, year of birth or year of observation. The fact that the patterns revealed are primarily vertical year-of-birth ones suggests that this is dominant over the year of observation.

4.9 Kernel smoothing gives us a more continuous form of the moving average. In fact, kernel smoothing can be used to fill in missing values; contrast Figure 9 with the moving-average results in Figure 5 and their missing triangles. The kernel smoother has automatically filled in the area of missing data with sensible values. This approach can be extended; by treating the future as a set of missing values, the kernel-smoothing approach can be made to give simple projections, as in Figure 10.

4.10 However, just how stable are these projections? Figure 11 shows how the projected patterns of mortality improvement differ for three equally spaced years of birth. In each of the four panels, a different starting point



Source: Own calculations using GAD data

Figure 11. Smoothed male mortality improvements in England and Wales with projections for three birth cohorts using full data; year-on-year improvements for individual years, smoothed using the Nadaraya-Watson two-dimensional kernel smoother with the Gaussian kernel function and a bandwidth parameter $\lambda = 4.5$

has been used. The top left panel shows the projection which would have been obtained with the data available up to and including 1996, with projections from 1998 onwards. Successive panels have two years' further data. If this projection method were stable, we would see similar extrapolations. We do not, and this instability would have a major financial consequence. For example, using each projected improvement basis, we can calculate a 16-year temporary annuity for the 1931 cohort, starting from the triennial experience in 1999 - 2001. We can compare the weakest and strongest of the above four bases with a basis with no improvements, as in Table 1. The reason for using a temporary annuity is that the data are only given up to age 89, and we do not want to muddy the comparison by making further assumptions about mortality at ages where we have no data.

4.11 The basis for future mortality improvements is critical for the correct reserving for both annuities and pensions. What Table 1 suggests is that the instability of these projected improvements can be half as important

Table 1. Increase in temporary annuity factor over basis without future mortality improvements; male single-life temporary annuities for 16 years from age 74, population mortality of 2003

Interest rate p.a.	Projection from 1998	2000	2002	2004
0%	2.8%	3.9%	2.5%	3.5%
3%	2.6%	3.5%	2.2%	3.1%
6%	2.3%	3.1%	2.0%	2.7%
9%	2.0%	2.8%	1.8%	2.4%

again, even at relatively high interest rates. Reserve increases of this magnitude are significant for both annuity portfolios and pension schemes. Table 1 shows, not only the financial significance of future improvements, but also the financial significance of the uncertainty surrounding those future improvements. We will return to the subject of the financial impact of such uncertainty in the next section. The topic of uncertainty in mortality projections is explored in greater detail in CMIB (2005).

4.12 Kernel smoothing improves on moving averages, because it can give a more continuous fitted model. We can also get projections of a sort. Furthermore, although the smoothing here was done 'by eye', there is a formal framework in kernel smoothing to select the optimal bandwidth parameter. However, kernel smoothing is still not a proper statistical model of the underlying stochastic process, and we would prefer such a statistical framework for both testing the fit and producing confidence intervals.

5. Estimating Year-of-Birth Effects using Penalised Splines

5.1 The smoothing methods which we have presented so far, moving averages and kernel smoothing, make no attempt to build a statistical model of the observed numbers of deaths. We compute the raw mortality improvements $\Delta m_{x,t}$, and then treat these values as the 'data' for our smoothing methods. The raw mortality rates are computed from very large amounts of data, so ignoring the different exposures which give rise to the different $\Delta m_{x,t}$ is unlikely to cause difficulties. However, the lack of a model causes difficulties for: (a) the calculation of confidence intervals; and (b) the choice of the level of smoothing. Our level of smoothing is chosen 'by eye', and this gives pleasing graphs, but how do we really know whether we have over-smoothed, i.e. removed some genuine features; or under-smoothed, i.e. left in some feature that is consistent with random variation?

5.2 We let $D_{x,t}$ be the random variable denoting the number of deaths at age x and year of birth t. We assume that $D_{x,t}$ has a Poisson distribution with mean $E^c_{x,t}\mu_{x,t}$, where $E^c_{x,t}$ is the central exposed to risk and $\mu_{x,t}$ is the force of mortality. If $d_{x,t}$ is the observed value of $D_{x,t}$, then the raw force of mortality is

$\hat{\mu}_{x,t} = d_{x,t}/E^c_{x,t}$. We must smooth the $\hat{\mu}_{x,t}$. Once smooth values of $\hat{\mu}_{x,t}$ are obtained we can convert to smooth values of $\hat{m}_{x,t}$, and hence to smooth mortality improvement rates.

5.3 We illustrate smoothing with penalised *B*-splines by considering some data on male assured lives at age 65; these data were supplied by the CMIB for years of observation from 1947 to 1999. It is convenient to model the force of mortality on the log scale (since the ratio of mortality at older ages to mortality at younger ages is very large). Our model is:

$$D_t \sim \mathcal{P}(E^c_t \mu_t) \quad \log \mu_t = \sum_{k=1}^{K} B_k(t)\theta_k \tag{1}$$

where the notation is as above, but where $x = 65$ and has been omitted for ease of presentation. This is a regression-type model, with the set of *B*-splines $\{B_1(t), \ldots, B_K(t)\}$ providing the regression basis (instead of the more familiar powers of t in a traditional polynomial regression). The regression coefficients are denoted by $\theta_1, \ldots, \theta_K$. The left panel of Figure 12 shows a single cubic *B*-spline. A cubic *B*-spline consists of cubic polynomial pieces bolted together at points known as knots; in the diagram the knots are equally spaced in calendar time at 1957.4, 1967.8, 1978.2, 1988.6 and 1999.0, and the *B*-spline is zero to the left of 1957.4 and to the right of 1999. The *B*-spline pieces are continuous, and have continuous first and second derivatives at the join points, shown $\circ$ in the left panel of Figure 12. The right panel in Figure 12 shows a basis of *B*-splines with $K = 8$. See de Boor (2001) for some actuarial references to smoothing mortality data with splines.

5.4 This model can be fitted with standard software, since the Poisson distribution together with the linear structure for $\log \mu_t$ defines a generalised linear model (see McCullagh & Nelder, 1989; Renshaw, 1991); the regression coefficients θ_k are chosen by maximum likelihood. The left panel in Figure 13 shows the result of fitting the regression with a basis of $K = 23$ *B*-splines. It is clear that if we had been smoothing by eye then we would not have been satisfied with the fit. The problem is that we have too many *B*-splines in our basis, and the resulting fit seems too flexible. In the 1970s and 1980s much effort went into the determination of the optimal number of *B*-splines, i.e. the number of *B*-splines that provides an optimal level of smoothing. Eilers & Marx (1996) proposed a different strategy. In Figure 13 the regression coefficients θ_k are plotted at the maximum value of $B_k(t)$. Eilers & Marx observed that the sort of under-smoothing evident in the left panel of Figure 13 (the saw-tooth effect in the $\circ$ plot) was a result of the erratic behaviour of the θ_k, and they proposed penalising this erratic behaviour by placing a difference penalty on adjacent θ_k, as in:

$$P(\theta) = (\theta_1 - 2\theta_2 + \theta_3)^2 + \ldots + (\theta_{K-2} - 2\theta_{K-1} + \theta_K)^2. \tag{2}$$

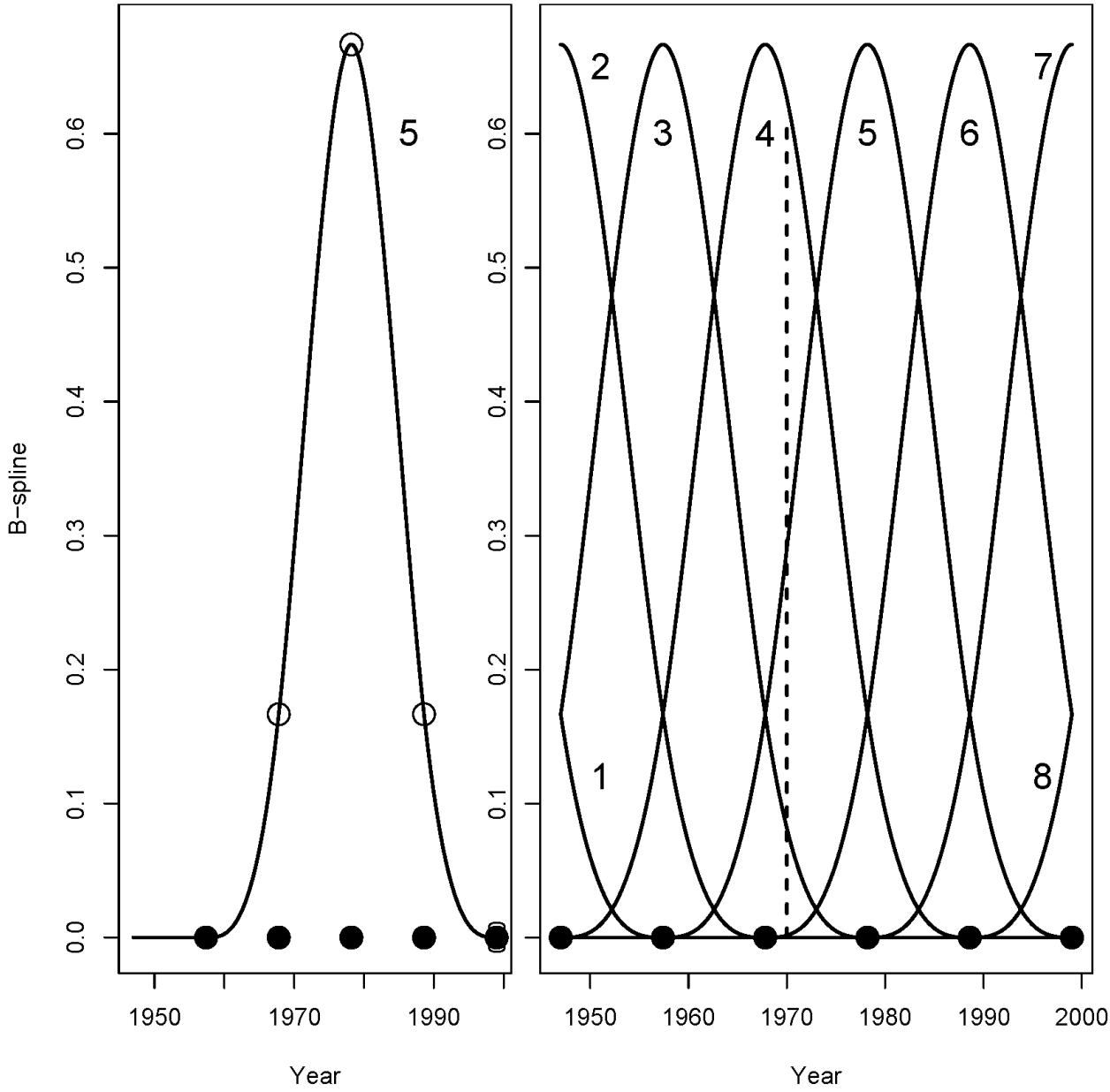


Figure 12. Left panel: a single cubic B-spline, $B_5(t)$, with knot positions, • and smooth joints, ○; right panel: a basis of $K = 8$ cubic B-splines with knot positions; a set of equally spaced knots is used for all splines in the basis, with each B-spline being zero-valued except on a finite interval

5.5 This defines a quadratic penalty; linear and cubic penalty functions are also possible. The penalty function is incorporated into the log-likelihood function $L(\theta)$, to give the penalised log-likelihood function $PL(\theta)$:

$$PL(\theta) = L(\theta) - \frac{1}{2}\lambda P(\theta). \tag{3}$$

5.6 This method is known as penalised B-spline regression, or P-splines

Figure 13. Smooth assured lives mortality with regression coefficients, ○; left panel: *B*-spline regression; right panel: *P*-spline regression (*B*-spline with structure penalty on regression coefficients)

for short. The parameter λ is the tuning constant, and plays exactly the same role as the bandwidth parameter of the kernel method. As with kernel smoothing, the larger the value of λ the stronger the smoothing. For a given value of λ, the regression coefficients are chosen by maximising $PL(\theta)$. One advantage of our statistical model is that we can use some statistical criterion to select the tuning constant λ; possibilities include the Akaike Information Criterion (AIC) (Akaike, 1987), the Bayesian Information Criterion (BIC) (Schwartz, 1978) or generalised cross-validation (GCV) (Craven & Wahba, 1979). These criteria balance (a) the closeness of fit of the observations to the fitted values with (b) the complexity of the fitted model.

5.7 The right panel of Figure 13 shows the result of smoothing with $K = 23$ cubic *B*-splines in the regression basis, but this time a quadratic penalty is used to smooth the regression coefficients; the tuning constant was chosen with BIC ($\lambda = 3900$). We need less smoothing with fewer *B*-splines in the basis. For example, with $K = 13$ we find $\lambda = 310$; the resulting fit is

indistinguishable from the right panel of Figure 13, and is omitted. Thus, the use of the BIC adjusts automatically for the number of *B*-splines in the basis by choosing the appropriate level of smoothing.

5.8 The method of *P*-splines has some similarities with moving averages and kernel smoothing. In view of Figure 13, we can interpret the coefficients θ_k as pseudo-observations; fitted values at year t are weighted averages of these pseudo-observations, where the weights are equal to the values of the non-zero *B*-splines at year t. For example, with the basis in the right panel of Figure 12, the estimate of $\log \mu_{1970}$ is:

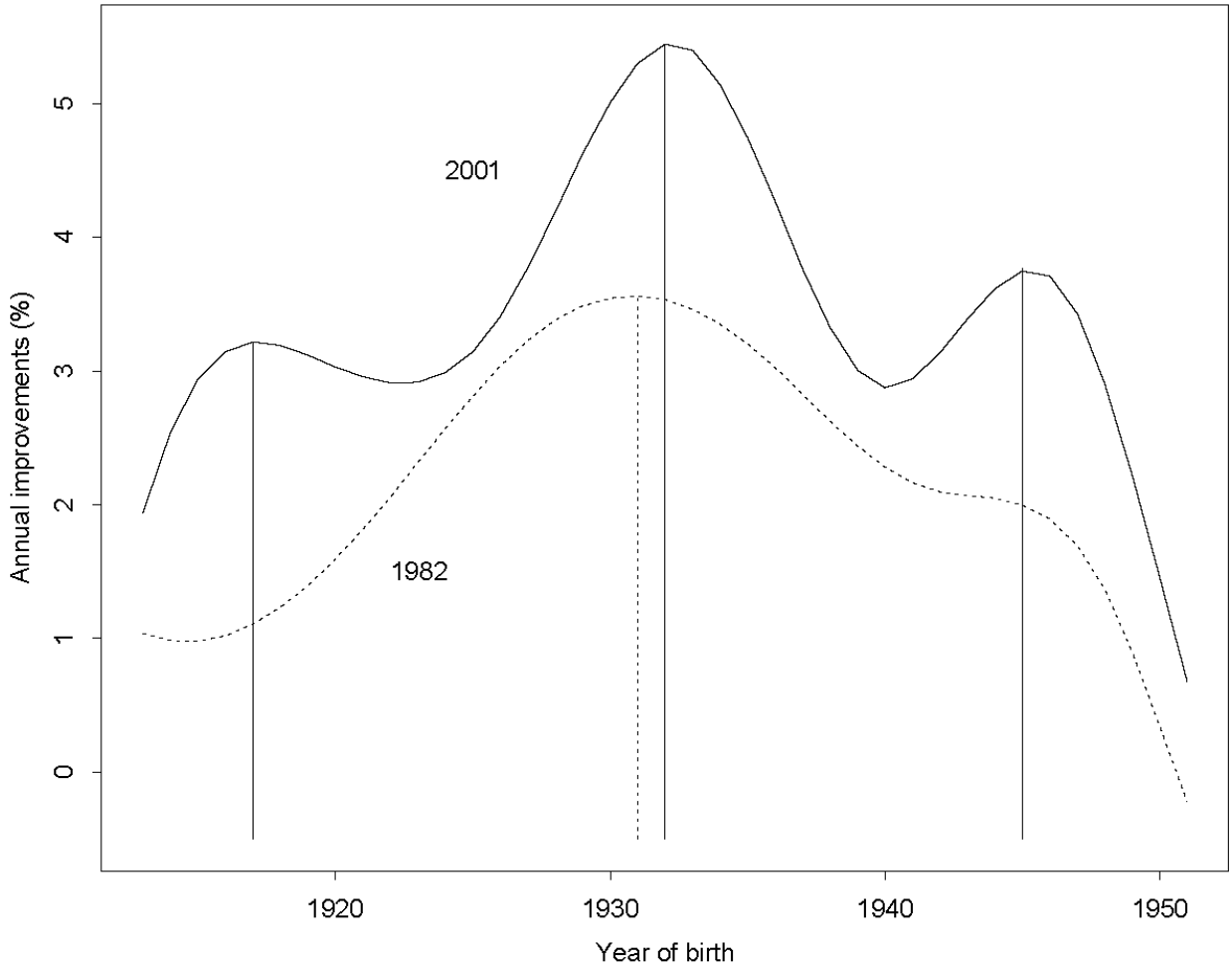
$$0.0817 \times \hat{\theta}_3 + 0.6267 \times \hat{\theta}_4 + 0.2901 \times \hat{\theta}_5 + 0.0016 \times \hat{\theta}_6 \tag{4}$$

where these weights are given by the intersection of the dashed line in Figure 12 with the labelled *B*-splines $B_3(t)$, $B_4(t)$, $B_5(t)$ and $B_6(t)$. Notice that at most four of the weights are non-zero at any year, and that the weights in equation 4 sum to unity. Thus, the *B*-spline weights move across the years from 1947 to 1999 in much the same way as the weights in moving average or kernel smoothing.

5.9 We use *P*-splines to estimate the triennium-to-triennium mortality improvements from 1980 - 82 to 1981 - 83 and from 1999 - 2001 to 2000 - 02. We estimated μ_t for each triennium from the deaths and the (central) exposures aggregated over the triennium. We then converted to mortality rates, q_t rates, and hence to mortality improvement rates. Figure 14 shows the results. The use of *P*-splines has resulted in much greater smoothing than moving averages gave in Figure 4, an example of a general property of smoothing methods which, like *P*-splines, choose the level of smoothing by balancing the fit to data with the complexity of the fitted curve. The main cohort effect for those born around 1931 is again identified. However, the subsidiary effects for the 1917 and 1945 cohorts are only found for the 2001 triennium.

5.10 We now turn to the problem of modelling mortality in terms of both age x and a second dimension t. We leave open, for the moment, whether to use calendar year or year of birth in t, as this is a model decision which we will illustrate later. For the purpose of presentation here, we will assume that t represents year of birth. The method of *P*-splines can be extended to cover the two-dimensional case as follows. First, we construct a pair of one-dimensional cubic *B*-spline bases, one for age, $\{B_1^a(x), \ldots, B_L^a(x)\}$, and one for year of birth $\{B_1^y(t), \ldots, B_K^y(t)\}$, as in Figure 12. Then, corresponding to equation 1, our model is:

$$D_{x,t} \sim \mathcal{P}(E_{x,t}^c \, \mu_{x,t}) \quad \log \mu_{x,t} = \sum_{l=1}^{L} \sum_{k=1}^{K} B_l^a(x) B_k^y(t) \theta_{l,k} \tag{5}$$



Source: Own calculations using GAD data

Figure 14. Male mortality improvement rates by year of birth in the U.K. for two triennia centred on 1982 and 2001 (*P*-spline smoothing)

where $D_{x,t}$ is the number of deaths and $E^c_{x,t}$ is the exposure at age x and year of birth t. There are a number of helpful analogies with one-dimensional *B*-splines and with moving averages. Just as Figure 12 shows the K functions in a one-dimensional *B*-spline basis, Figure 15 shows the KL functions in a (small) two-dimensional basis. In practice, we would have a large number of such hills; for example, with $L = 20$ and $K = 10$ we would have 200 overlapping hills, and these provide a flexible basis for two-dimensional regression.

5.11 The KL regression coefficients can be thought of as pseudo-observations located at the peaks of the hills, with the non-zero $B^a_l(x)B^y_k(t)$ in the vicinity of (x, t) providing the appropriate weights. Finally, we note that the product form of the weights in equation 5 corresponds with the product form of the moving average weights which we used in ¶3.9.

5.12 Equation 5 defines a generalised linear model in exactly the same

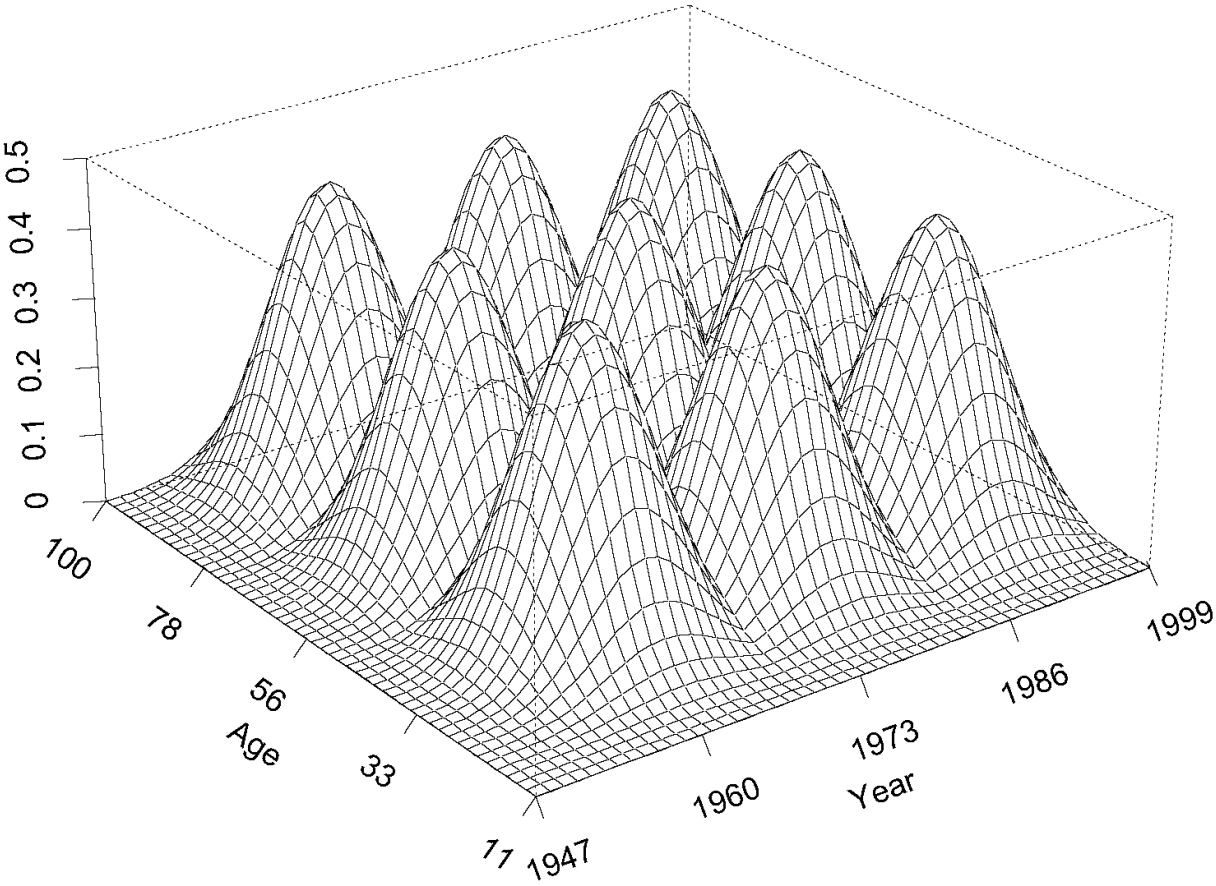


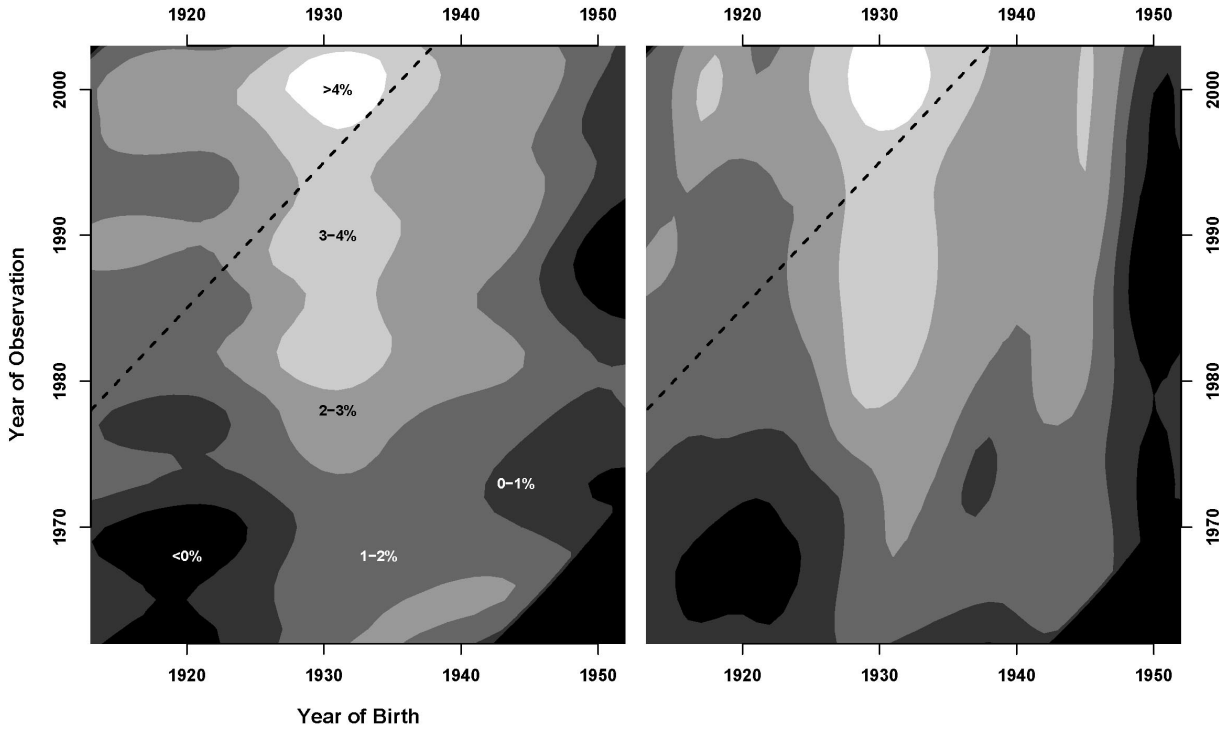
Figure 15. An example two-dimensional *B*-spline basis

way as equation 1, and so can be fitted with standard software. Similarly, the resulting fit will suffer from the same non-smoothness as is evident in the left panel of Figure 13. We force smoothness on the fitted surface by penalising the regression coefficients along both the age and the year-of-birth axes. The penalised log likelihood has the form:

$$PL(\theta) = L(\theta) - \frac{1}{2}\lambda_a P_a(\theta) - \frac{1}{2}\lambda_y P_y(\theta) \tag{6}$$

where there are now two tuning constants: λ_a in age and λ_y in year of birth. As in the one-dimensional case, the regression coefficients are chosen by maximising equation 6 for given joint values of the tuning constants λ_a and λ_y; the values of these tuning constants are, in turn, chosen by minimising BIC. We refer to this model with the data classified by age of death and year of birth as the age-cohort model.

5.13 We can also apply two-dimensional *P*-spline smoothing to mortality data classified by age and by year of observation (the age-period model), but it is important to realise that this is not equivalent to the age-



Source: Own calculations using GAD data

Figure 16. Smoothed male mortality improvements in England and Wales using penalised splines; the left panel shows the results of a model based on age and year of observation (age-period, BIC = 10569.97); the right panel shows the results of a better fitting alternative model based on age and year of birth (age-cohort, BIC = 10469.84); a cohort-period model was also considered, but gave a much poorer fit, due to the strength of the age signal

cohort model. In the age-cohort model, the penalties are applied along the age and year-of-birth axes, whereas in the age-time model they are applied along the age and year-of-observation axes. Figure 16 shows the result of applying the method to the GAD data classified in both ways. The cohort effect centred on 1931 is clearly seen in both panels; the right panel also suggests a second cohort effect centred on 1945. The BIC criterion can also be used to choose between different classes of models, and here the age-cohort model is preferred to the age-time model. We conclude that year of birth gives a stronger signal than year of observation in a model of mortality. Certainly, the age-cohort model is more effective at showing the presence of cohort effects. The dominance of cohort effects over period effects echoes the findings of Kermack *et al.* (1934).

5.14 Figure 16 can be contrasted with Figure 9. In the terminology of kernel smoothing, the P-spline model has selected the optimum bandwidth by a selection criterion (in this case optimising the Bayesian information criterion). The resulting fitted model optimally balances broader year-of-

birth patterns against sharper, more pronounced effects for certain individual years of birth. The goodness of fit of these models is demonstrated by local inspection, such as demonstrated in the right panel of Figure 13.

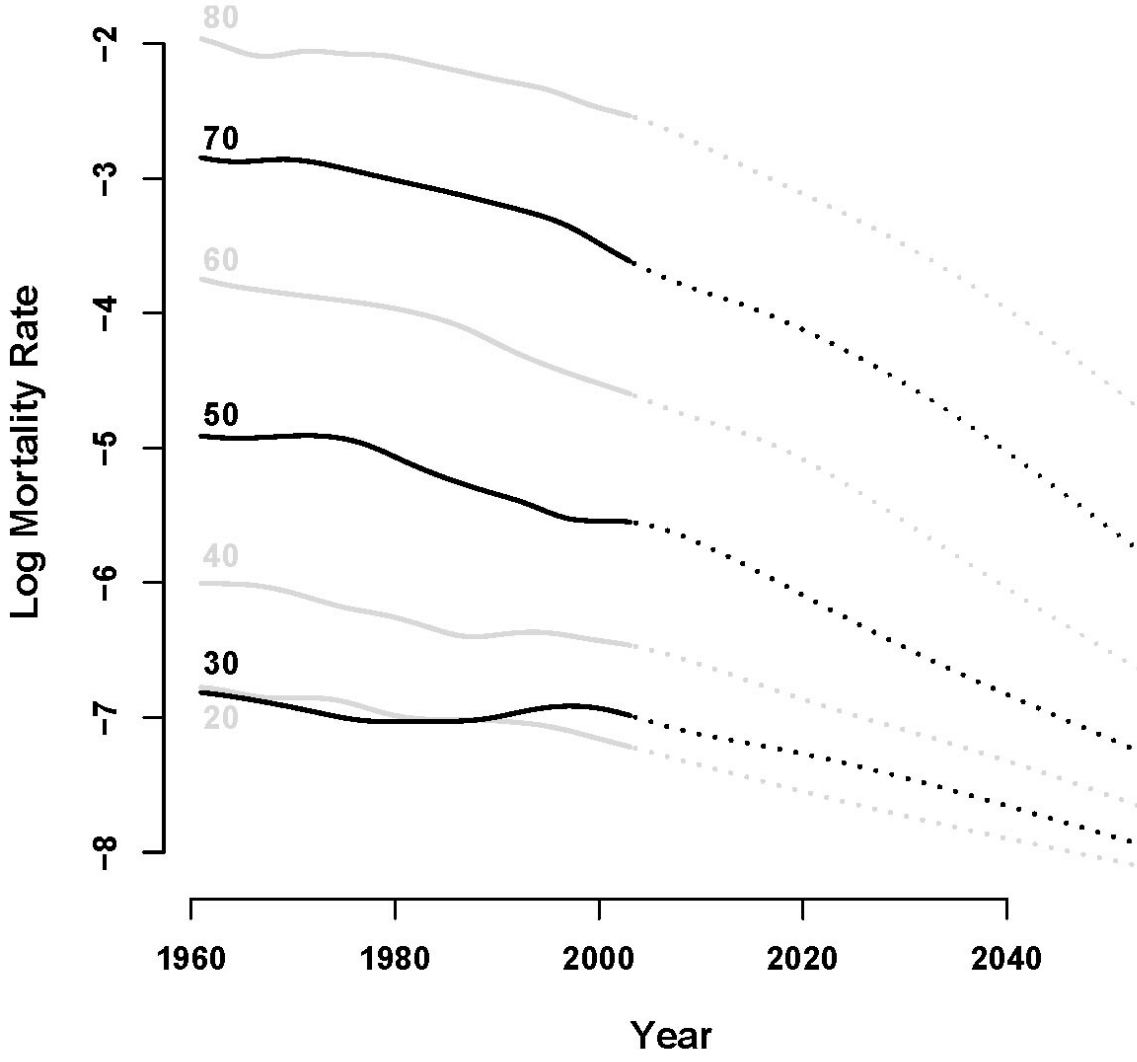
6. Forecasting Mortality using Penalised Splines

6.1 Forecasting is a natural consequence of the penalty function. We sketch the argument in one dimension. Suppose we want to forecast log mortality at age 65, as shown in the right panel of Figure 13, up to, say, 2050, say. We obtained Figure 13 by defining a basis of $K = 23$ cubic B-splines over the period 1947 to 1999. To forecast, we extend the basis to cover the years up to 2050; this gives a basis with $K = 43$. In equation 3, the log-likelihood function $L(\theta)$ remains the same as before, since we have no additional data; in particular, $L(\theta)$ depends only on $\theta_1, \ldots, \theta_{23}$. However, the penalty function $P(\theta)$ contains all 43 coefficients. We will maximise the forecast $PL(\theta)$ by choosing the same values of $\theta_1, \ldots, \theta_{23}$ as before, if the value of the penalty $P(\theta)$ is unaltered; this is achieved by linear extrapolation on θ_{22} and θ_{23}, since the additional penalty in equation 2 is zero, as is readily checked.

6.2 It is important to appreciate the role of the penalty function in forecasting. We have used a quadratic penalty in equation 2, and this leads to linear extrapolation. Different choices are possible, and this choice has little effect on the fit to data. However, the effect on the forecast is dramatic. With a linear penalty, the forecast log mortality is constant, while, with a cubic penalty, it is quadratic. Thus, the choice of penalty function corresponds to a decision on the future course of mortality. A linear penality would project no further mortality improvements, which is not consistent with past trends. We feel that the quadratic penalty, and hence a linear forecast on a log scale, sits comfortably with Figure 13 and similar plots for other ages.

6.3 We use the same method in two dimensions. We extend the one-dimensional basis for the calendar year to 2050, say, and this, in turn, extends the two-dimensional basis; Figure 15 helps to visualise the new basis. We then maximise the penalised log-likelihood in equation 6. For any given age the extrapolations are no longer exactly linear on a log scale, since the age penalty tends to maintain the age structure across ages. This tension between the age and year penalties results in a consistent, two-dimensional forecast of the whole mortality table. Figure 17 shows the projected mortality rates at selected ages which result from this.

6.4 Ages between 70 and 80 are key for annuity pricing and reserving, including both deferred and immediate annuities. Deferred annuities, in particular, are heavily dependent on long-term mortality projections. A portfolio of deferred annuities will often include ages as young as 30. The mortality projections in Figure 17 have mortality at age 70 falling by three-



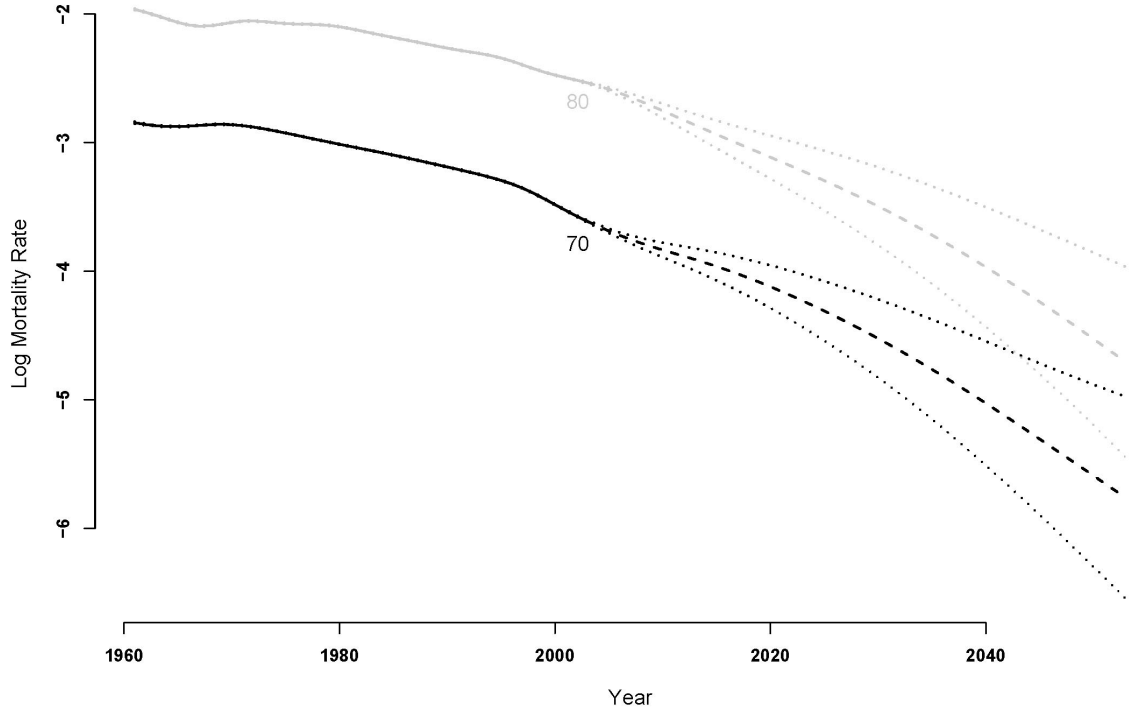
Source: Own calculations using GAD data

Figure 17. Logarithm of smoothed mortality rates for male experience in England and Wales with projections for key ages using penalised splines

quarters over the next 40 years, an equivalent average mortality improvement of around $3\frac{1}{2}$%p.a.

6.5 However, deferred annuities do not have quite as long an average term as this; the amounts-weighted average age in a portfolio of deferred annuities is usually higher, mainly due to the age-related value of accrued benefits. Nevertheless, even relatively short-term projections for deferred annuities are eye opening; the projections above have mortality at age 80 falling by more than half over the next 20 years, again an equivalent average mortality improvement of $3\frac{1}{2}$%p.a. If you prefer, the above projection has 80-year-olds experiencing the same mortality rates in 2040 as 60-year-olds did in the mid 1970s.

6.6 Thus, some of the greatest implications of continued cohort mortality



Source: Own calculations using GAD data

Figure 18. Logarithm of smoothed mortality rates for male experience in England and Wales with projections for ages 70 and 80 using penalised splines and 95% confidence intervals

improvements arise for deferred annuities, be they on life assurers' balance sheets or, more commonly, in defined benefit pension schemes. Of course, no projection is worth anything without some measure of the uncertainty surrounding it, and Figure 18 shows the projections at ages 70 and 80, together with 95% confidence intervals. As can be seen, not only do these ages have very strong projected mortality improvements, but considerable uncertainty as well. Note that these confidence intervals critically assume that the assumed model is the appropriate one. What they do not allow for is model risk, which is to say that the choice of a given model is itself subject to the risk that there are other, possibly more appropriate, possibly yet unknown models. As such, these confidence intervals should be viewed as guides to uncertainty appropriate to these projections, and not absolutist statements that mortality trajectories are 95% certain to lie within them. The topic of model risk in the context of mortality projections is discussed extensively in CMIB (2005).

7. Conclusion

7.1 Moving averages are fine for initial exploratory work in mortality improvements. However, greater insights into mortality improvements can be gained from kernel smoothing and P-splines. Both kernel smoothing and

P-splines give intermediate smoothed values, i.e. a smoothed continuous surface is created. Also, in contrast to moving averages, both kernel smoothing and P-splines can cope easily with missing data points. Since future values are, by definition, 'missing data points', both methods can therefore provide projections of existing trends by simply treating the future as 'missing data'.

7.2 A key further benefit of P-splines is that the smoothing results from a statistical fitting procedure. This procedure not only gives a statistical test of badness of fit, but it also enables confidence intervals to be placed on the future projections. This formal statistical framework enables the testing of the relative strength of effects due to year of observation and year of birth. Applied to this data set, we find that male mortality improvements amongst the population of England and Wales are more strongly driven by year-of-birth effects than they are by year-of-observation effects.

7.3 We have shown that a model for mortality patterns can be made to fit better by using year of birth as a rating factor, and we have shown why year of birth can be a useful combined proxy for both early-life conditions and lifetime effects. There are other, more significant rating factors, but the penalised-spline approach can be extended to include these other rating factors if the data are available. Examples of this extension can be found in Currie *et al.* (2004b).

Acknowledgments

The authors thank Adrian Gallop and the Government Actuary's Department (GAD) for providing the data and useful suggestions. The authors thank the late Professor Sir Richard Doll and Dr Jillian Boreham for providing data from their smoking survey. Finally, the authors thank David Robinson, Keith Miller and Richard Willets and two anonymous scrutineers for their helpful comments.

The work of I. D. Currie was supported by a grant from the Actuarial Research Fund. The work of J. G. Kirkby was supported by an Engineering and Physical Sciences Research Council studentship, with further support from the CMIB.

Data manipulation was carried out using simple C programs and R, an open-source statistics package available free of charge at
http://www.r-project.org

References

The following list includes, not only the works referred to in the paper, but other publications that might prove helpful by way of further background.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrica*, **52**, 317-333.

Ames, B. N. (1998). Micro-nutrients prevent cancer and delay ageing. *Toxicology Letters*, **102-103**, 5-18.

BENGTSSON, T. & LINDSTROM, M. (2000). Childhood misery and disease in later life. *Population Studies* (Camb.), **54(3)**, 263-277.

DE BOOR, C. (2001). A practical guide to splines (Revised edition). *Applied Mathematical Sciences*, **27**.

CMIB (CONTINUOUS MORTALITY INVESTIGATION BUREAU) (1999). *Report number 17.* Institute and Faculty of Actuaries.

CMIB (MORTALITY SUB-COMMITTEE) (2002). An interim basis for adjusting the 92 Series mortality projections for cohort effects. Working paper No. 1.

CMIB (MORTALITY SUB-COMMITTEE) (2005). Projecting future mortality: towards a proposal for a stochastic methodology. Working paper No. 15.

CRAVEN, P. & WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377-390.

CURRIE, I.D., DURBAN, M. & EILERS, P.H.C. (2003). Using P-splines to extrapolate two-dimensional Poisson data, *Proceedings of 18th International Workshop on Statistical Modelling*, Leuven, Belgium, 97-102.

CURRIE, I.D., DURBAN, M. & EILERS, P.H.C. (2004a). Array regression: an approach to smoothing data on arrays. Unpublished paper.

CURRIE, I.D., DURBAN, M. & EILERS, P.H.C. (2004b). Smoothing and forecasting mortality rates. *Statistical Modelling*, **4,** 279-298.

DOBLHAMMER, G. & VAUPEL, J.W. (2001). Lifespan depends on month of birth. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5, 2934-2939.

DOLL, R. & HILL, A.B. (1954). The mortality of doctors in relation to their smoking habits: a preliminary report. *British Medical Journal*, 1954, **ii**, 1451-1455.

DOLL, R., PETO, R., BOREHAM, J. & SUTHERLAND, I. (2004). Mortality in relation to smoking: 50 years' observations on male British doctors. *British Medical Journal*, 2004, **328**, 1519-.

DURBAN, M., CURRIE, I.D. & EILERS, P.H.C. (2002). Using P-splines to smooth two-dimensional Poisson data. *Proceedings of 17th International Workshop on Statistical Modelling*, Chania, Crete, 207-214.

EILERS, P.H.C. & MARX, B.D. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science*, **11**, 89-121.

FINCH, C.E. & CRIMMINS, E.M. (2004). Inflammatory exposure and historical changes in human life-spans. *Science*, 17 September 2004, **305**.

FOREY, B.A., LEE, P.N. & FRY, J.S. (1993). Updating U.K. estimates of age, sex and period specific cumulative constant tar cigarette consumption per adult. *Thorax*, **53**, 875-878.

GAD (GOVERNMENT ACTUARY'S DEPARTMENT) (2003). Interim life tables. http://www.gad.gov.uk

GAVRILOV, L.A., GAVRILOVA, N.S., SEMENOVA, V.G., EVDOKUSHKINA, G.N., KRUT'KO, V.N., GAVRILOVA, A.L., EVDOKUSHKINA, N.N. & LAPSHIN, E.V. (1997). Maternal age and lifespan of offspring. *Doklady Akademii Nauk*, **354**, 4.

GAVRILOV, L.A. & GAVRILOVA, N.S. (1999). Season of birth and human longevity. *Journal of Anti-Aging Medicine*, **2**(4), 365-366.

GAVRILOV, L.A. & GAVRILOVA, N.S. (2000). Human longevity and parental age at conception. In (Robine *et al.*, eds.) *Sex and longevity: sexuality, gender, reproduction, parenthood.* Springer-Verlag, Berlin, Heidelberg.

GAVRILOV, L.A. & GAVRILOVA, N.S. (2001). The reliability theory of aging and longevity. *Journal of Theoretical Biology*, **213**, 527-545.

GAVRILOVA, N.S., GAVRILOV, L.A., EVDOKUSHKINA, G.N. & SEMYONOVA, V.G. (2002). Early life conditions and later sex differences in adult lifespan. Paper presented at 2002 annual meeting of Population Association of America, May 9-11 2002, Atlanta.

GAVRILOV, L.A. & GAVRILOVA, N.S. (2003). Early-life factors modulating lifespan, In (Rattan, S.I.S., ed.) *Modulating aging and longevity.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

GAVRILOV, L.A. & GAVRILOVA, N.S. (2004). Early-life programming of aging and longevity:

the idea of high initial damage load (the HIDL hypothesis). *Annals of the New York Academy of Sciences*, **1019**, 496-501.

GLUCKMAN, P.D. & HANSON, M.A. (2004). Living with the past: evolution, development, and patterns of disease. *Science*, 17 September 2004, **305**.

IZUMOTO, Y., INOUE, S. & YASUDA, N. (1999). Schizophrenia and the influenza epidemics of 1957 in Japan. *Biological Psychiatry*, **46**, 1, 119-124.

KERMACK, W.O., MCKENDRICK, A.G. & MCKINLEY, P.L. (1934). Death rates in Britain and Sweden: some regularities and their significance. *Lancet*, 698-703. (Reprinted in *International Journal of Epidemiology*, 2001, **30**, 678-683).

LEE, P.N., FRY, J.S. & FOREY, B.A. (1990). Trends in lung cancer, chronic obstructive lung disease, and emphysema death rates for England and Wales 1941-85 and their relation to trends in cigarette smoking. *Thorax*, **45**, 657-665.

LIESTOL, K. (1981). A note on the influence of factors early in development on later reproductive function. *Annals of Human Biology*, **8**(6), 559-565.

MADJID, M., ABOSHADY, I., AWAN, I., LITOVSKY, S. & WARD CASSCELLS, S. (2004). Influenza and cardiovascular disease: is there a causal relationship? *Texas Heart Institute Journal*, **31**.

MCCARRON, P., OKASHA, M., MCEWAN, J. & SMITH, G.D. (2002). Height in young adulthood and risk of death from cardio-respiratory disease. *American Journal of Epidemiology*, **155**(8), 683-687.

MCCULLAGH, P. & NELDER, J.A. (1989). *Generalized linear models*, 2nd ed. Chapman and Hall, London.

NADARAYA, E.A. (1964). On estimating regression. *Theory of Probability and Applications*, **9**.

PELTONEN, M. & ASPLUND, K. (1996). Age-period-cohort effects on stroke mortality in Sweden from 1969-1993 and forecasts up to the year 2003. *Stroke*, **27**(11), 1981-1985.

R DEVELOPMENT CORE TEAM (2004). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.r-project.org

REINERT-AZAMBUJA, M.I. (2004). Spanish flu and early 20th-century expansion of a coronary heart disease-prone subpopulation. *Texas Heart Institute Journal*, **31**.

RENSHAW, A.E. (1991). Actuarial graduation practice and generalised linear and non-linear models. *Journal of the Institute of Actuaries*, **118**, 295-312.

RICHARDS, S.J. & JONES, G.L. (2004). Financial aspects of longevity risk. Paper presented to the Staple Inn Actuarial Society.

ROSEBLOOM, T.J., VAN DER MEULEN, J.H., OSMOND, C., BARKER, D.J., RAVELLI, A.C. & BLEKER, O.P. (2001). Adult survival after prenatal exposure to the Dutch famine of 1944-45. *Paediatric Perinatal Epidemiology*, **15**(3), 220-225.

SCHWARTZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 462-464.

SELTEN, J.P., BROWN, A.S., MOONS, K.G., SLAETS, J.P., SUSSER, E.S. & KAHN, R.S. (1999). Prenatal exposure to influenza as a risk factor for adult schizophrenia. *Schizophrenia Research*, **38**(2-3), 85-91.

STANNER, S.A., BULMER, K., ANDRES, C., LANTSEVA, O.E., BORODINA, V., POTEEN, V.V. & YUDKIN, J.S. (1997). Does malnutrition in utero determine diabetes and coronary heart disease in adulthood? Results from the Leningrad siege study. *British Medical Journal*, **315**, 1342-1348.

TODD, G.F., LEE, P.N. & WILSON, M.J. (1976). Cohort analysis of cigarette smoking and of mortality from four associated diseases. Occasional paper 3, Tobacco Research Council, London.

VIJAYAKUMAR, M., FALL, C.H., OSMOND, C. & BARKER, D.J. (1995). Birth weight, weight at one year, and left ventricular mass in adult life. *British Heart Journal*, **73**(4), 363-367.

WAND, M.P. & JONES, C.M. (1995). *Kernel smoothing*, Monographs on Statistics and Applied Probability, Chapman and Hall.

WATSON, G.S. (1964). Smooth regression analysis. *Sankhya*, Series A.

WILLETS, R.C. (1999). Mortality in the next millennium. Paper presented to the Staple Inn Actuarial Society.

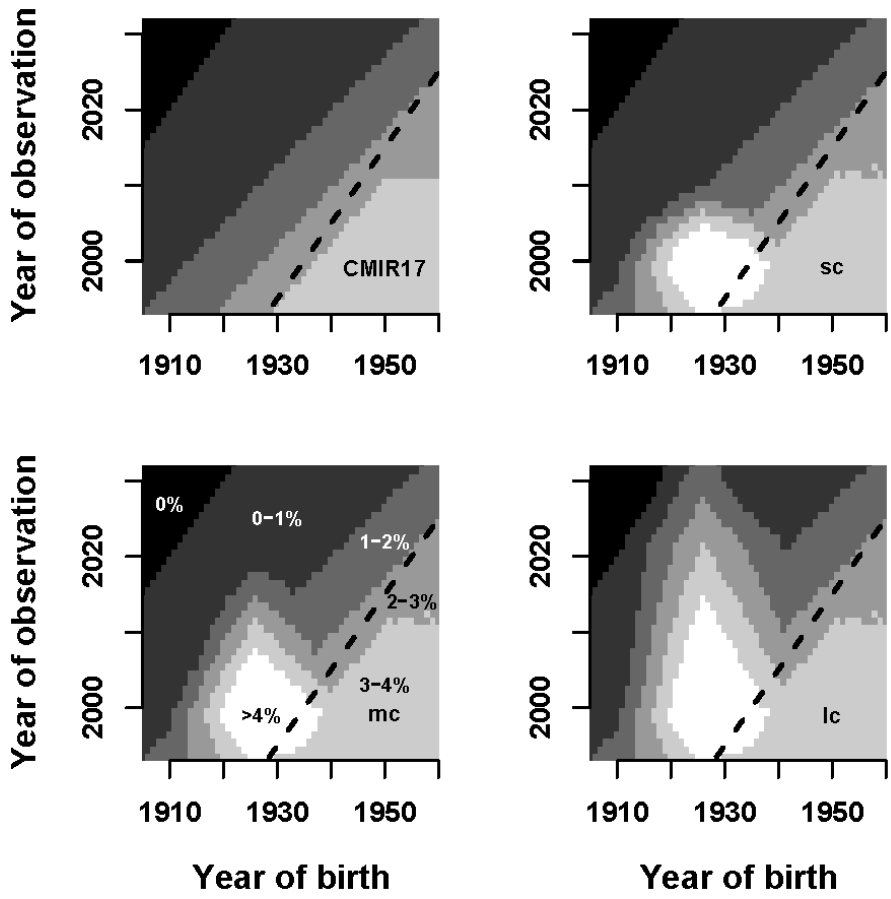
WILLETS, R.C. (2004). The cohort effect: insights and explanations. *British Actuarial Journal*, **10**, 833-877.

APPENDIX

PROJECTION BASES IN CURRENT ACTUARIAL USE

A.1 Figure 19 shows four projection bases in common use by U.K. life offices for annuity business at the time of writing. Most major U.K. life offices currently use the mid-intensity (or medium) cohort projection (bottom left panel), or some modification thereof, for reserving for male annuitants. Figure 19 shows how the CMIB Projections Working Party (2002) superimposed the three recent cohort projection bases on top of the older CMIR17 basis, which is clearly shown to be an age-based improvement basis only.

A.2 The CMIB makes no recommendation as to the suitability of these



Source: Own calculations using data in CMIB (1999, 2002)

Figure 19. Recent CMIB mortality improvement projections; in Z-layout from top left: original CMIR17 projection, then short, medium and long-cohort projections; the contour bands are labelled on the medium-cohort projection in the bottom left, the projection in most common use at the time of writing; the dashed line on each panel connects age 65

projections for any particular purpose, leaving it up to the individual actuary to decide. An actuary can make a case for not using these particular cohort projections for financial work related to annuities: (i) they are derived from lives-based, not amounts-based patterns of improvement; (ii) they are derived from male experience only; (iii) they are derived from the assured population (endowments), not the annuitant population; and (iv) they are derived from the experience of a relatively small population (assured lives are few above age 65). Nevertheless, as this paper shows, some sort of cohort-based mortality projection is required for longevity risks in the U.K. Figure 20 shows the mortality improvements implied by the P-spline model with age-cohort penalties in exactly the same format as Figure 19.

A.3 One thing which is not clear is the extent to which defined benefit pension schemes reserve adequately for future mortality improvements arising from the cohort effect. Without routine disclosure of mortality bases, it is impossible to be sure, but, in the words of Richards & Jones (2004): "some of the greatest future surprises from longevity risk may come from companies with large defined benefit pension schemes."

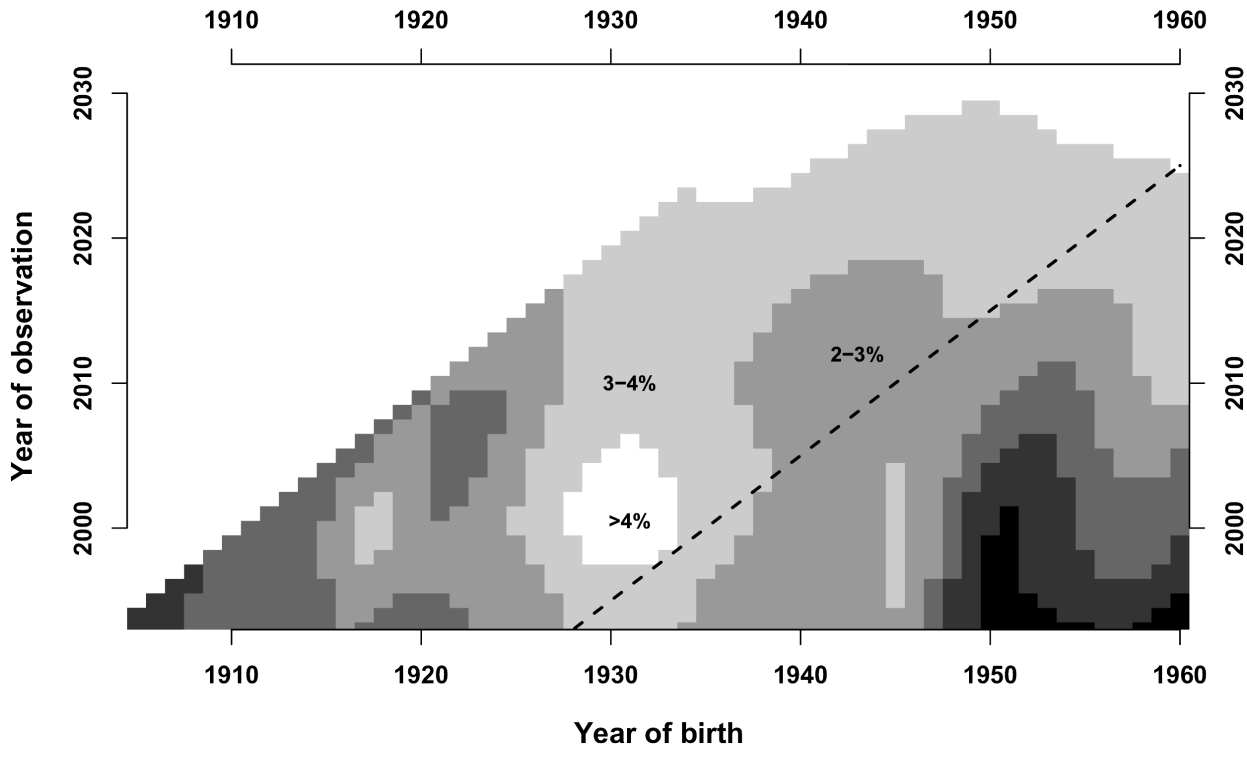


Figure 20. Mortality improvement projections according to P-spline model with penalties across age and cohort. The projection is based on the ONS data for the male population of England & Wales. The dashed line connect age 65