Backtesting Stochastic Mortality Models: An Ex Post Evaluation of Multiperiod-Ahead Density Forecasts

Kevin Dowd,* Andrew J. G. Cairns,† David Blake,‡ Guy D. Coughlan,§ David Epstein,** and Marwa Khalaf-Allah††

Abstract

This study sets out a backtesting framework applicable to the multiperiod-ahead forecasts from stochastic mortality models and uses it to evaluate the forecasting performance of six different stochastic mortality models applied to English & Welsh male mortality data. The models considered are the following: Lee-Carter's 1992 one-factor model; a version of Renshaw-Haberman's 2006 extension of the Lee-Carter model to allow for a cohort effect; the age-period-cohort model, which is a simplified version of Renshaw-Haberman; Cairns, Blake, and Dowd's 2006 two-factor model; and two generalized versions of the last named with an added cohort effect. For the data set used herein, the results from applying this methodology suggest that the models perform adequately by most backtests and that prediction intervals that incorporate parameter uncertainty are wider than those that do not. We also find little difference between the performances of five of the models, but the remaining model shows considerable forecast instability.

1. Introduction

Cairns et al. (2009) recently examined the empirical fits of eight different stochastic mortality models, variously labeled M1–M8. Seven[1] of these models were further analyzed in a second study, Cairns et al. (2008), which evaluated the ex ante plausibility of the models' probability density forecasts. Among other findings, that study found that one of these models—M8, a version of the Cairns-Blake-Dowd (CBD) model (Cairns et al. 2006) with an allowance for a cohort effect—generated implausible forecasts on U.S. data, and consequently this model was also dropped from further consideration. A third study, Dowd et al. (2010), then examined the ''goodness of fit'' of the remaining six models by analyzing the statistical properties of their various residual series. The six models that were examined are the following: M1, the Lee-Carter model (Lee and Carter 1992); M2B, a version of Renshaw and Haberman's

* Kevin Dowd, PhD, is Visiting Professor at the Pensions Institute, Cass Business School, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom, Kevin.Dowd@hotmail.co.uk.

† Andrew J. G. Cairns, PhD, is Professor of Financial Mathematics at the Maxwell Institute for Mathematical Sciences, and Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom, A.Cairns@ma.hw.ac.uk.

‡ David Blake, PhD, is Professor of Pension Economics and Director of the Pensions Institute at the Pensions Institute, Cass Business School, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom, E-d.blake@city.ac.uk.

§ Guy D. Coughlan, PhD, is Managing Director at the Pension Advisory Group, JPMorgan Chase Bank, 125 London Wall, London EC2Y 5AJ, United Kingdom, guy.coughlan@jpmorgan.com.

** David Epstein, PhD, is Executive Director at the Pension Advisory Group, JPMorgan Chase Bank, 125 London Wall, London EC2Y 5AJ, United Kingdom, david.uk.epstein@jpmorgan.com.

†† Marwa Khalaf-Allah, PhD, is Vice President at the Pension Advisory Group, JPMorgan Chase Bank, 125 London Wall, London EC2Y 5AJ, United Kingdom, marwa.khalafallah@jpmorgan.com.

[1] The model omitted from this second study was the P-splines model of Currie et al. (2004). This model was dropped in part because of its poor performance relative to the other models when assessed by the Bayes Information Criterion, and in part because it cannot be used to project stochastic mortality outcomes.

cohort-effect generalization of the Lee-Carter model (Renshaw and Haberman 2006); M3B, a version of the age-period-cohort model (e.g., Osmond 1985; Jacobsen et al. 2002), which is a simplified version of M2B;[2] M5, the CBD model; and M6 and M7, two alternative cohort-effect generalizations of the CBD model. Details of these models' specifications are given in Appendix A.

It is quite possible for a model to provide a good in-sample fit to historical data *and* produce forecasts that appear to be ''plausible'' ex ante, but still produce poor ex post forecasts, that is, forecasts that differ significantly from subsequently realized outcomes. A ''good'' model should therefore produce forecasts that perform well out-of-sample when evaluated using appropriate forecast evaluation or backtesting methods, *as well as* provide good fits to the historical data and plausible forecasts ex ante. The primary purpose of the present paper is, accordingly, to set out a backtesting framework that can be used to evaluate the ex post forecasting performance of a mortality model.

A secondary purpose of the paper is to illustrate this backtesting framework by applying it to the six models listed above. The backtesting framework is applied to each model over various forecast horizons using a particular data set, namely, LifeMetrics data for the mortality rates of English & Welsh males[3] for ages 60–84 and spanning the years 1961–2008.[4]

The backtesting of mortality models is still in its infancy. Most studies that assess mortality forecasts focus on ex post forecast errors (e.g., Keilman 1997, 1998; National Research Council 2000; Koissi et al. 2006). This is limited insofar as it ignores all information contained in the probability density forecasts, except for the information reflected in the mean forecast or ''best estimate.'' A more sophisticated treatment—and a good indicator of the current state of best practice in the evaluation of mortality models—is provided by Lee and Miller (2001). They evaluated the performance of the Lee-Carter (1992) model by examining the behavior of forecast errors (comparing MAD, RMSE, etc.) and plots of ''percentile error distributions,'' although they did not report any formal test results based on these latter plots. However, they also reported plots showing mortality prediction intervals and subsequently realized mortality rates, and the frequencies with which realized mortality observations fell within the prediction intervals. These provide a more formal (i.e., probabilistic) sense of model performance. More recently, Continuous Mortality Investigation (2006) included backtesting evaluations of the P-spline model; they (2007) also included backtesting evaluations of M1 and M2. Their evaluations were based on plots of realized outcomes against forecasts, where the forecasts included both projections of central values and projections of prediction intervals, which also give some probabilistic sense of model performance.

The backtesting framework used in this paper can best be understood if we outline the following key steps:

1. We begin by selecting the metric of interest, namely, the forecasted variable that is the focus of the backtest. Possible metrics include the mortality rate, life expectancy, future survival rates, and prices of annuities and other life-contingent financial instruments. Different metrics are relevant for different purposes: for example, in evaluating the effectiveness of a hedge of longevity or mortality risk, the relevant metric is the monetary value of the underlying exposure. In this paper, we focus on the mortality rate itself, but, in principle, backtests could be conducted on any of these other metrics as well.
2. We select the historical ''lookback'' window, which is used to estimate the parameters of each model for any given year: thus, if we wish to estimate the parameters for year t and we use a lookback

[2] M2B and M3B are the versions of M2 and M3 that assume an ARIMA(1,1,0) process for the cohort effect (Cairns et al. 2008).

[3] See Coughlan et al. (2010) and www.lifemetrics.com for the data and a description of LifeMetrics. The original source of the data was the U.K. Office for National Statistics. Note that we derive mortality rates directly from deaths and exposures data and not from graduated q rates. The latter series is not appropriate because it has been smoothed, and this smoothing may hinder reliable modeling, in particular of the cohort effect.

[4] We should note that when estimating the parameters of the cohort effect, we excluded cohorts for which there were less than 5 observations.

window of length n, then we are estimating the parameters for year t, using observations from years $t - n$ to $t - 1$. In this paper, we use a fixed-length[5] lookback window of 20 years.[6]

3. We select the horizon (i.e., the ''lookforward'' window) over which we will make our forecasts, based on the estimated parameters of the model. In the present study, we focus on relatively long-horizon forecasts, because it is with the accuracy of these forecasts that pension plans are principally concerned, but that also pose the greatest modeling challenges.
4. Given the above, we decide on the backtest to be implemented and specify what constitutes a ''pass'' or ''fail'' result. Note that we use the term ''backtest'' here to refer to any method of evaluating forecasts against subsequently realized outcomes. ''Backtests'' in this sense might involve the use of plots whose goodness of fit is interpreted informally, as well as formal statistical tests of predictions generated under the null hypothesis that a model produces adequate forecasts.

The above framework for backtesting stochastic mortality models is a very general one.

Within this broad framework, we implement the following four backtest procedures for each model, noting that the metric (mortality rate) and lookback windows (20 years) are the same in each case:

- *Contracting horizon backtests:* First, we evaluate the convergence properties of the forecast mortality rate to the actual mortality rate at a specified future date. We chose 2008 as the forecast date and we examine forecasts of that year's mortality rate, where the forecasts are made at different dates in the past. The first forecast was made with a model estimated from 20 years of historical observations up to 1980, the second with the model estimated from 20 years of historical observations up to 1981, and so forth, up to 2008. In other words, we are examining forecasts with a fixed end date and a contracting horizon from 28 years ahead down to one year ahead. The backtest procedure then involves a graphical comparison of the evolution of the forecasts toward the actual mortality rate for 2008, and intuitively ''good'' forecasts should converge in a fairly steady manner toward the realized value for 2008.
- *Expanding horizon backtests:* Second, we consider the accuracy of forecasts of mortality rates over expanding horizons from a common fixed start date, or ''stepping off'' year. For example, the start date might be 1980, and the forecasts might be for one up to 28 years ahead, that is, for 1981 up to 2008. The backtest procedure again involves a graphical comparison of forecasts against realized outcomes, and the ''goodness'' of the forecasts can be assessed in terms of their closeness to the realized outcomes.
- *Rolling fixed-length horizon backtests:* Third, we consider the accuracy of forecasts over fixed-length horizons as the stepping off date moves sequentially forward through time. This involves examining plots of mortality prediction intervals for some fixed-length horizon (e.g., 20 years) rolling forward over time—with subsequently realized outcomes superimposed on them.
- *Mortality probability density forecast tests:* Fourth, we carry out formal hypothesis tests in which we use each model as of one or more start dates to simulate the forecasted mortality probability density at the end of some horizon period, and we then compare the realized mortality rate(s) against this forecasted probability density (or densities). The forecast ''passes'' the test if each realized rate lies within the more central region of the forecast probability density, and the forecast ''fails'' if it lies too far out in either tail to be plausible given the forecasted probability density.

[5] The use of a fixed-length lookback window simplifies the underlying statistics and means that our results are equally responsive to the ''news'' in any given year's observations. By contrast, an expanding lookback window is more difficult to handle and also means that, as time goes by, the ''news'' contained in the most recent observations receives less weight, relative to the expanding number of older observations in the lookback sample.

[6] We have chosen a lookback window of 20 years to estimate the parameters of the model. The choice of a 20-year window is a subjective one based on the limited length of our dataset (of 49 years) and the need to strike a balance between having a lookback window long enough to get ''reasonable'' results, on the one hand, and the need to have enough observations left over to accommodate ''reasonably'' long forecast horizons, on the other.

For each of the four classes of test, we examine two types of mortality forecast. The first of these are forecasts in which it is assumed that the parameters of the mortality models are known with certainty. The second are forecasts, in which we make allowance for the fact that the parameters of the models are only estimated, that is, their ''true'' values are unknown. We would regard the latter as more plausible, and evidence from other studies suggests that allowing for parameter uncertainty can make a considerable difference to estimates of quantifiable uncertainty (e.g., Cairns et al. 2006; Dowd et al. 2006; Blake et al. 2008). For convenience we label these the ''parameter certain'' (PC) and ''parameter uncertain'' (PU) cases, respectively.[7] More details of the latter case are given in Appendix B.

This paper is organized as follows. Section 2 considers the contracting horizon backtests, Section 3 considers the expanding horizon backtests, Section 4 considers the rolling fixed-length backtests for an illustrative horizon of 15 years, and Section 5 considers the mortality probability forecast tests. Section 6 concludes.

2. CONTRACTING HORIZON BACKTESTS: EXAMINING THE CONVERGENCE OF FORECASTS THROUGH TIME

The first kind of backtest in our framework examines the consistency of forecasts for a fixed future year (in our example below, the year 2008) made in earlier years. For a well-behaved model we would expect consecutive forecasts to converge toward the realized outcome as the date the forecasts are made (the stepping-off date) approaches the forecast year.

Figures 1 and 2 show plots of forecasts of the 2008 mortality rate for 65-year-old and 84-year-old males, respectively, made in years 1980, 1981, . . . , 2007. The central lines in each plot are the relevant model's median forecast of the 2008 mortality rate based on 5,000 simulation paths, and the dotted lines on either side are estimates of the model's 90% prediction interval or risk bounds (i.e., 5th and 95th percentiles). The starred point in each chart is the realized mortality rate for that age in 2008.

These plots show the following patterns:

- For models M1, M3B, M5, M6, and M7, the forecasts converge and the prediction intervals narrow in a fairly stable way over time toward a value close to the realized value.[8] These results reflect that the underlying state variables exhibit a smooth progression from one sample period to the next and, with one exception, have parameter values that are plausible and obey permitted ranges.[9]
- For these same models, the PU prediction intervals are notably wider than the PC prediction intervals, but negligible difference is seen between the PC and PU median plots.
- Also for these same models, the age 65 forecasts tend to show a downward trend, suggesting a bias that increases with the forecast horizon, whereas the age 84 forecasts show little or no obvious bias.
- For M2B, the forecasts are clearly unstable, exhibiting major spikes: these projections reflect estimates of the cohort state variables that are sometimes very unstable and highly implausible as we move from one sample to the next; these, in turn, lead to estimates of the parameters of the gamma process that are also sometimes very unstable and implausible, and well outside permitted ranges.

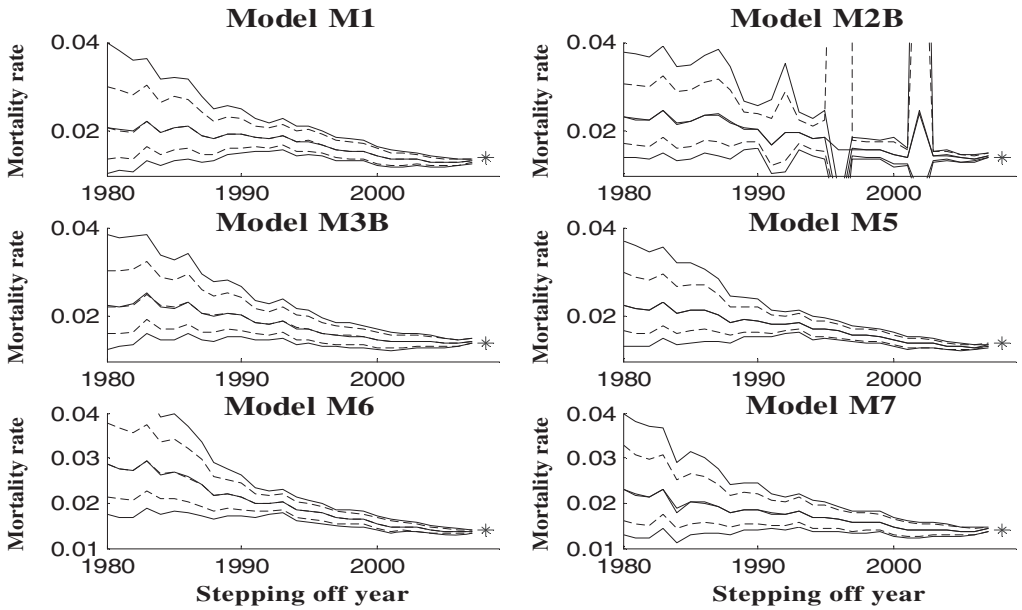
[7] To be more precise, the PU versions of the model use a Bayesian approach to take account of the uncertainty in the parameters driving the period and, where applicable, the cohort effects.

[8] There is, however, the exception of the age 84 M1 projections in the early 1980s (see Fig. 2), which decline almost to a point in 1984 and then start to fan out again. This behavior is explained by the fact that the early 1980s $\beta^{(2)}$ estimates fall sharply to values close to zero, and, for this model, it is these parameters that determine the dispersion of the fan charts.

[9] The one exception relates to the sample years 1977–1996 estimate of the $\alpha^{(\gamma)}$ parameter for model M7, which at 1.011 is just over the permitted range of $(-1,+1)$, and this makes it impossible for our code to simulate random values of this parameter from a posterior distribution. To remedy this problem, this $\alpha^{(\gamma)}$ value was reset to 0.98 for the projections in Figures 1 and 2.

Figure 1

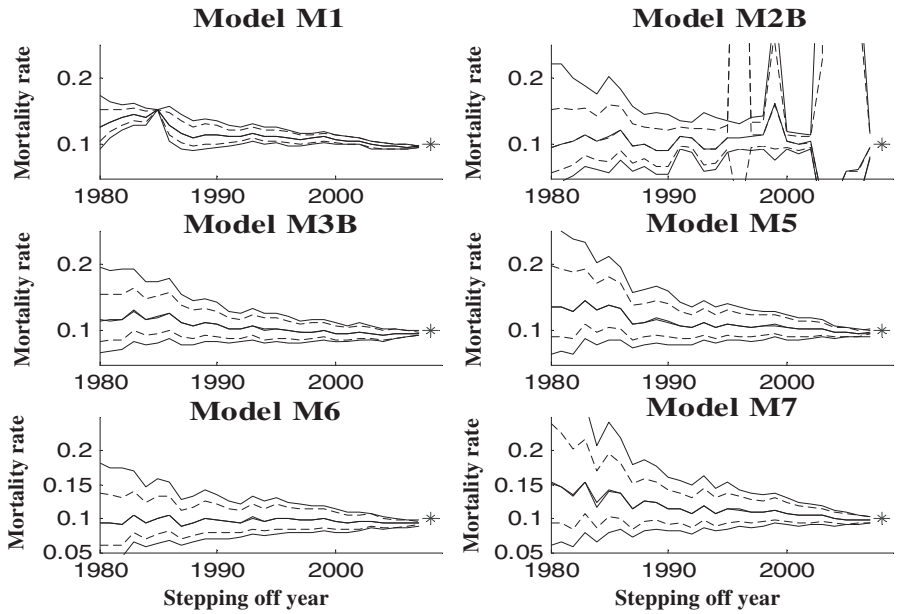
Forecasts of the 2008 Mortality Rate from 1980 Onward: Males Aged 65



Notes: Forecasts based on estimates using English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window. The stepping-off year is the final year in the rolling window. The fitted model is then used to estimate the median and 90% prediction interval for both parameter-certain forecasts (given by the dashed lines) and parameter-uncertain cases (given by the solid lines). The realized mortality rate for 2008 is denoted by an asterisk. Based on 5,000 simulation trials.

Figure 2

Forecasts of the 2008 Mortality Rate from 1980 Onward: Males Aged 84



Notes: Forecasts based on estimates using English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window. The stepping-off year is the final year in the rolling window. The fitted model is then used to estimate the median and 90% prediction interval for both parameter-certain forecasts (given by the dashed lines) and parameter-uncertain cases (given by the solid lines). The realized mortality rate for 2008 is denoted by an asterisk. Based on 5,000 simulation trials.

Note that we have evaluated the convergence of the models only for one end date and one data set. This is insufficient to draw general conclusions about the forecasting capabilities of the various models, but the instability observed in M2 (or at least our variant of M2, M2B) is an issue.

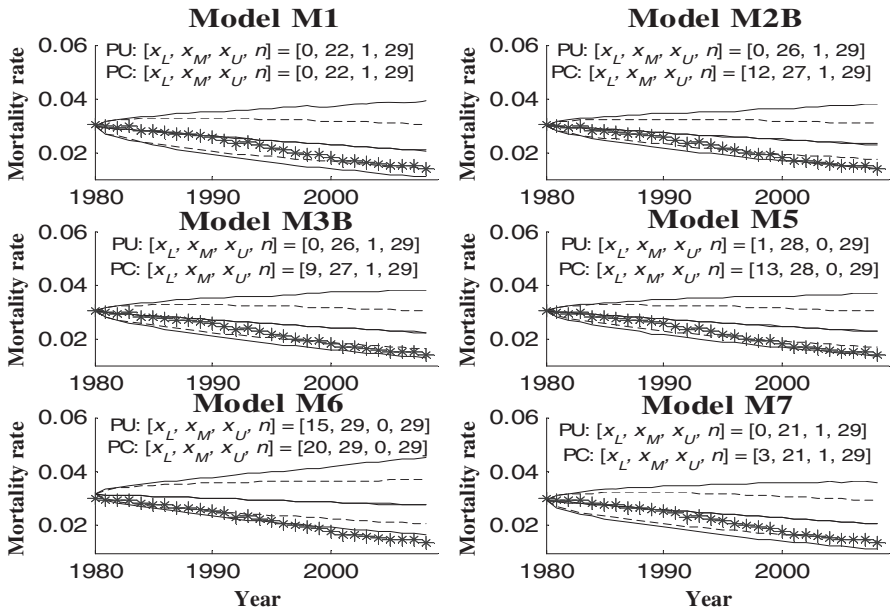
3. EXPANDING HORIZON BACKTESTS

In the second class of backtest, we consider the accuracy of forecasts over increasing horizons against the realized outcomes for those horizons. Accuracy is reflected in the degree of consistency between the outcome and the prediction interval associated with each forecast.

These backtests are best evaluated graphically using charts of mortality prediction intervals. Figure 3 shows the mortality prediction intervals for age 65 for forecasts starting in 1980 with model parameters estimated using data from the preceding 20 years, and Figure 4 shows the same for age 84. The charts show the 90% prediction intervals (or ''risk bounds'') as dashed lines for the PC forecasts and continuous lines for the PU forecasts. Roughly speaking, if a model is adequate, we can be 90% confident of any given outcome occurring between the dashed risk bounds if we ''believe'' the PC forecasts, and we can be 90% confident of any given outcome occurring between the continuous risk bounds if we ''believe'' the PU forecasts. The figures also show the forecasted median mortality forecasts as dotted lines for the PC forecasts and continuous lines for the PU forecasts (although these are in fact usually hard to distinguish), and superimposed on the figures are the realized outcomes indicated by stars.

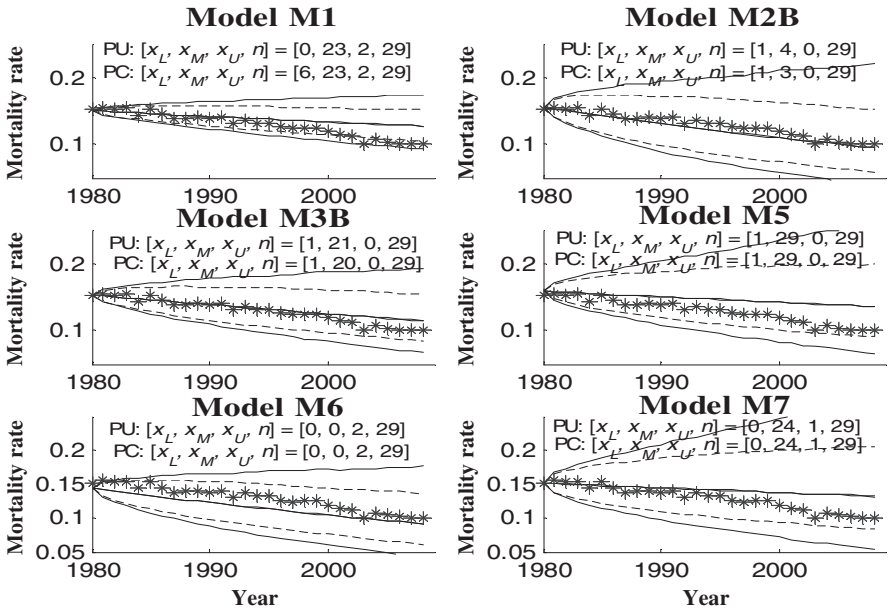
The figures also show quadruplets in the form $[x_L, x_M, x_U, n]$, where n is the number of mortality forecasts, x_L is the number of lower exceedances or the number of realized outcomes out of n falling below the lower risk bound, x_M is the number of observations falling below the projected median, and

Figure 3
Mortality Prediction-Interval Charts from 1980: Males Aged 65



Notes: Forecasts based on estimates using English & Welsh male mortality data for ages 60–84 and years 1961–1980. The dashed lines refer to the forecast medians and bounds of the 90% prediction interval for the parameter-certain (PC) forecasts, and solid lines are their equivalents for the parameter-uncertain (PU) forecasts. The realized mortality rates are denoted by an asterisk. For each of these cases, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

Figure 4
Mortality Prediction-Interval Charts from 1980: Males Aged 84



Notes: Forecasts based on estimates using English & Welsh male mortality data for ages 60–84 and years 1961–1980. The dashed lines refer to the forecast medians and bounds of the 90% prediction interval for the parameter-certain (PC) forecasts, and solid lines are their equivalents for the parameter-uncertain (PU) forecasts. The realized mortality rates are denoted by an asterisk. For each of these cases, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

x_U is the number of upper exceedances or observations falling above the upper risk bound.[10] These statistics provide useful indicators of the adequacy of model forecasts. If the forecasts are adequate, for any given PC or PU case, we would expect x_L and x_U to be about 5%, and we would expect x_M to be about 50%. Too many exceedances, on the other hand, would suggest that the relevant bounds were incorrectly forecasted: for example, too many realizations above the upper risk bound would suggest that the upper risk bound is too low, too many observations below the median bound would suggest that the forecasts are biased upwards, and too many observations below the upper risk bound would suggest that the lower risk bound is too high.

The results in these charts can be summarized as follows:

- The prediction intervals all show the same basic shape: they fan out somewhat over time around a gradually decreasing trend and show a little more uncertainty on the upper side than on the lower side.
- The PU risk bounds are notably wider than their PC equivalents, and virtually no differences are seen between the PC- and PU-based median projections.
- For age 65, the realized values tend to gravitate toward the lower bounds of the prediction intervals. For most models, a considerable number of lower bound violations exist for the PC projections, but (with the exception of M6) few such violations for the PU projections. By this criterion, the PU projections perform better than the PC ones.
- For age 84, the realized values are usually much closer to the median projections than is the case for age 65, and with the exception of the M1 projections, we find almost no lower bound violations.

[10] Note, however, that the positions of the individual observations within the prediction intervals are not independent. If, for example, the 1990 observation is ''low,'' then the 1995 observation is also likely to be ''low.''

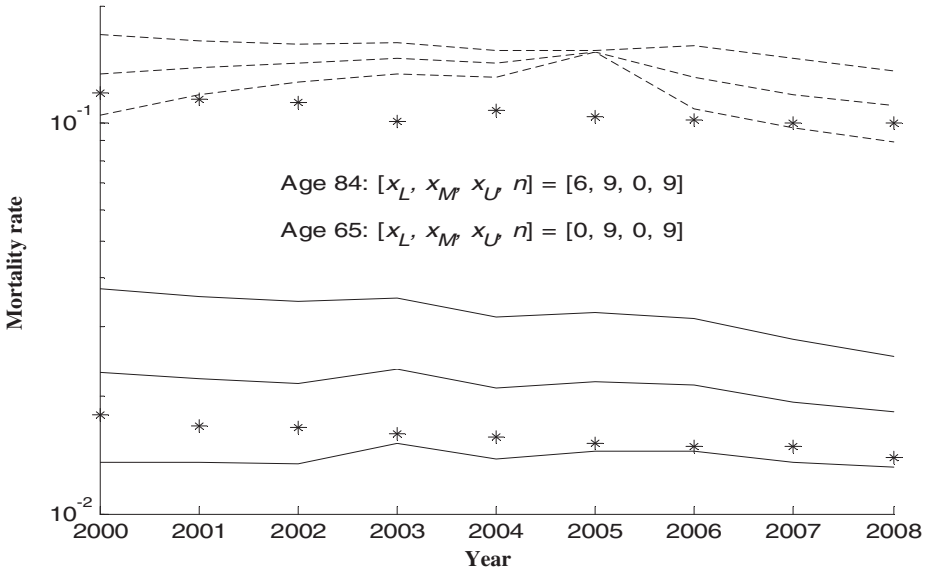
- The results of these last two bullet points suggest an upward bias in the age 65 projections but much less bias in the age 84 projections.

4. ROLLING FIXED-LENGTH HORIZON BACKTESTS

In the third class of backtests, we consider the accuracy of forecasts over fixed horizon periods as these roll forward through time. Once again, accuracy is reflected in the degree of consistency between the collective set of realized outcomes and the prediction intervals associated with the forecasts. We examine the case of an illustrative 20-year forecast horizon, and hereafter we restrict ourselves to backtesting the performance of the PU forecasts only. Figures 5–10 show the rolling prediction intervals for each of the six models in turn. Each plot shows the rolling prediction intervals and realized mortality outcomes for ages 65 and 84 over the years 2000–2008, where the forecasts are obtained using data from 20 years or more before. Figure 5 gives the prediction intervals for model M1, Figure 6 gives the rolling prediction intervals for M2B, and so forth. To facilitate comparison, the *y*-axes in all the charts have the same range running from 0.01 to 0.2 on a logarithmic scale.

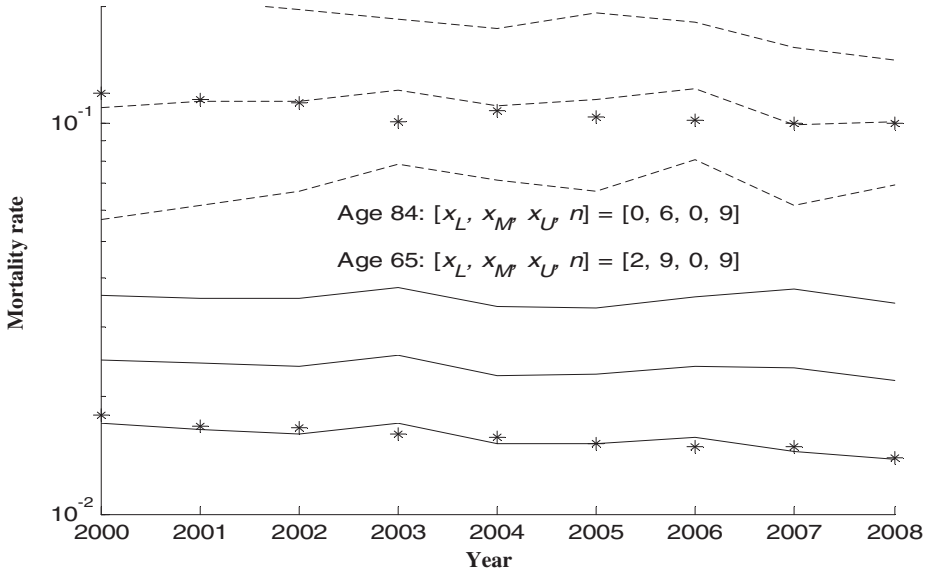
These charts indicate that there is not much to choose between the other models. We also find that the age 65 forecasts have a notable upward bias with the realized values often close to or below the lower bounds, whereas the age 84 forecast show only a very slight upward bias.

Figure 5
Rolling 20-Year-Ahead Prediction Intervals: M1



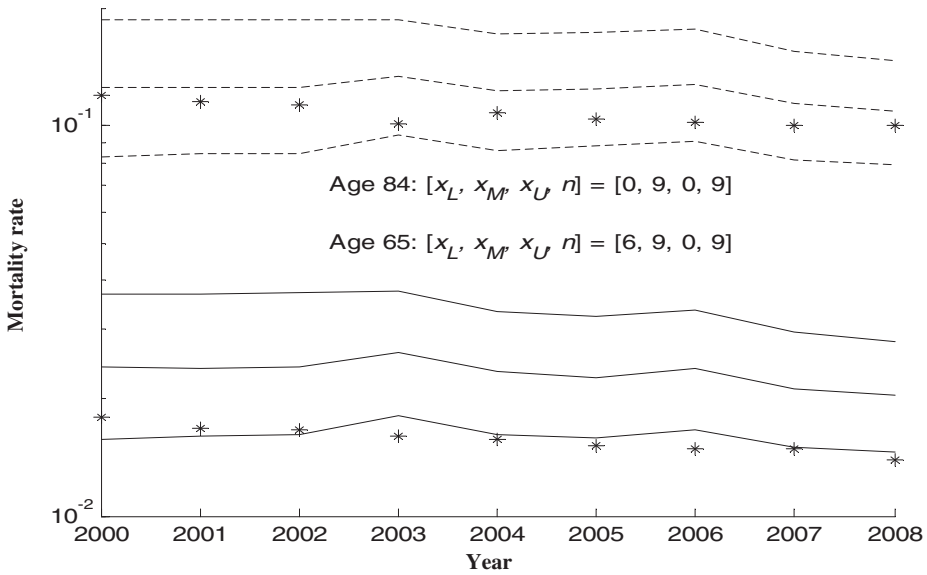
Notes: Model estimates based on English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window starting from 1961–1980. The continuous and dashed lines are the parameter-uncertain (PU) forecast medians and bounds of the 90% prediction interval for ages 65 and 84, respectively. The realized mortality rates are denoted by an asterisk. For each age, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

Figure 6
Rolling 20-Year-Ahead Prediction Intervals: M2B



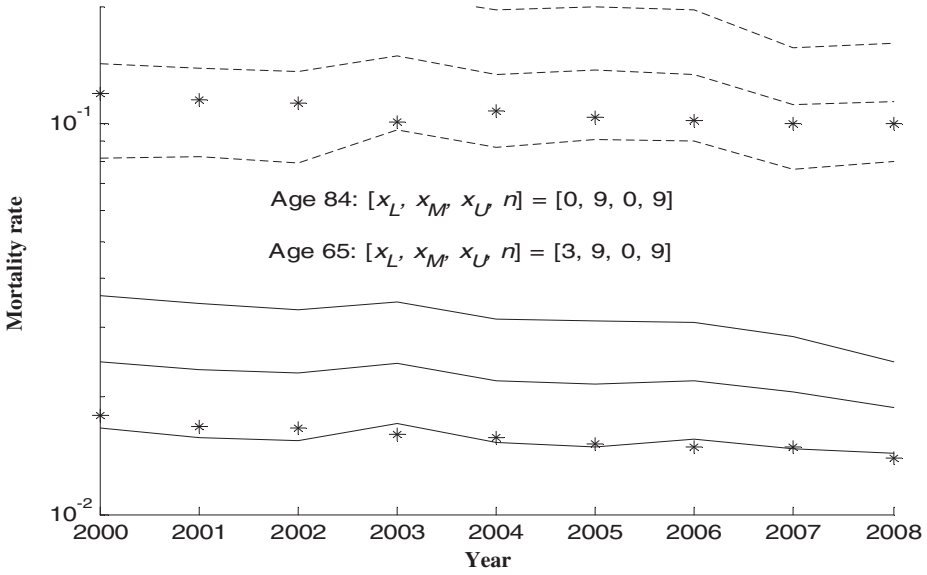
Notes: Model estimates based on English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window starting from 1961–1980. The continuous and dashed lines are the parameter-uncertain (PU) forecast medians and bounds of the 90% prediction interval for ages 65 and 84, respectively. The realized mortality rates are denoted by an asterisk. For each age, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

Figure 7
Rolling 20-Year-Ahead Prediction Intervals: M3B



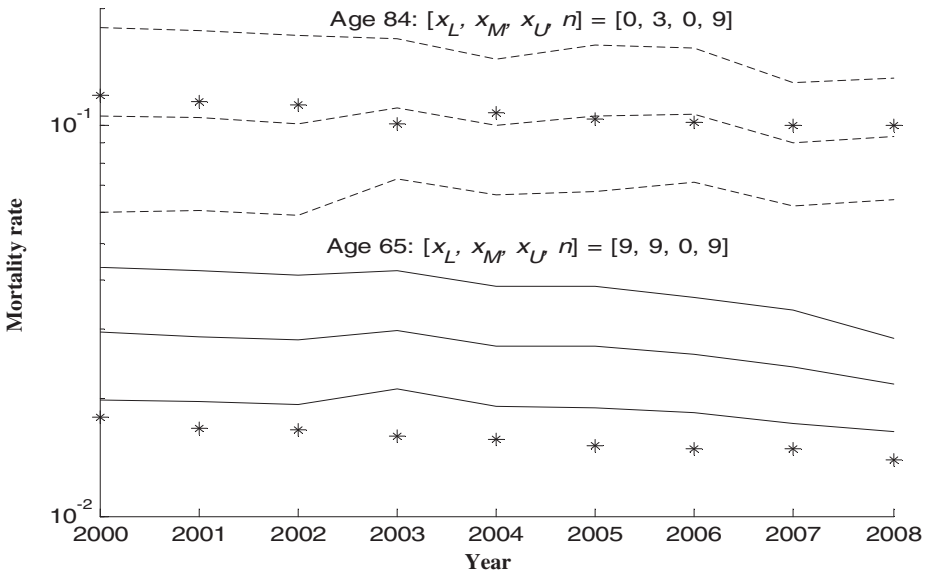
Notes: Model estimates based on English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window starting from 1961–1980. The continuous and dashed lines are the parameter-uncertain (PU) forecast medians and bounds of the 90% prediction interval for ages 65 and 84, respectively. The realized mortality rates are denoted by an asterisk. For each age, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

Figure 8
Rolling 20-Year-Ahead Prediction Intervals: M5



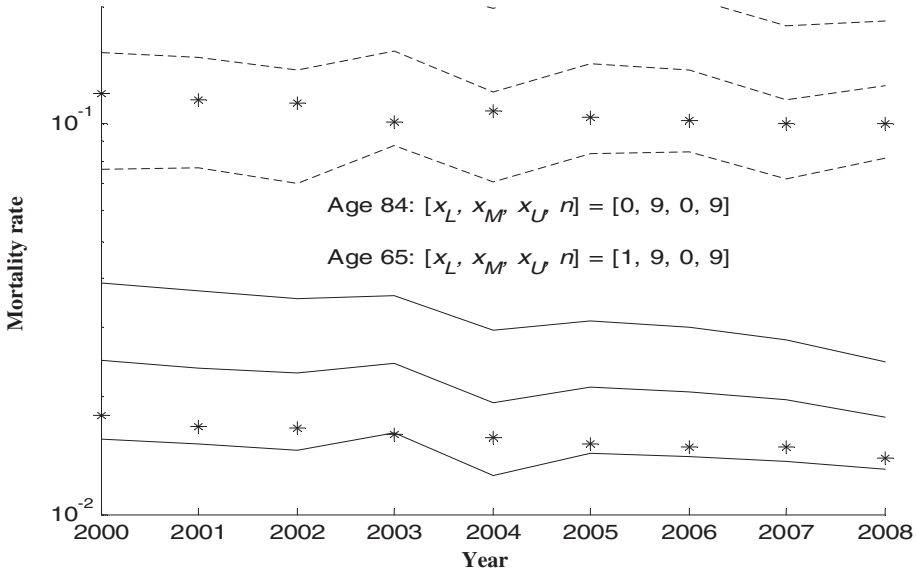
Notes: Model estimates based on English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window starting from 1961–1980. The continuous and dashed lines are the parameter-uncertain (PU) forecast medians and bounds of the 90% prediction interval for ages 65 and 84, respectively. The realized mortality rates are denoted by an asterisk. For each age, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

Figure 9
Rolling 20-Year-Ahead Prediction Intervals: M6



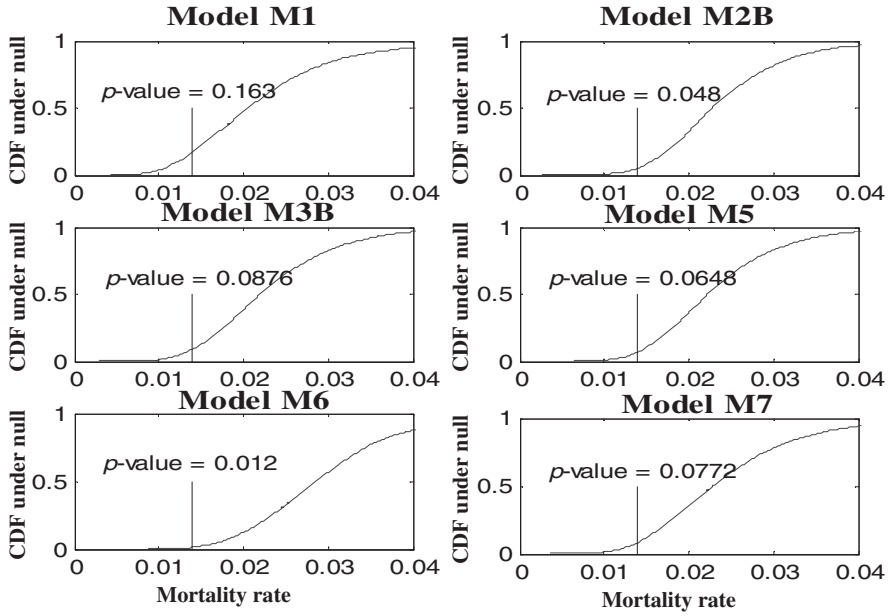
Notes: Model estimates based on English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window starting from 1961–1980. The continuous and dashed lines are the parameter-uncertain (PU) forecast medians and bounds of the 90% prediction interval for ages 65 and 84, respectively. The realized mortality rates are denoted by an asterisk. For each age, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

Figure 10
Rolling 20-Year-Ahead Prediction Intervals: M7



Notes: Model estimates based on English & Welsh male mortality data for ages 60–84 and a rolling 20-year historical window starting from 1961–1980. The continuous and dashed lines are the parameter-uncertain (PU) forecast medians and bounds of the 90% prediction interval for ages 65 and 84, respectively. The realized mortality rates are denoted by an asterisk. For each age, x_L and x_M are the numbers of realized rates below the lower 5% and 50% prediction bounds, x_U is the number of realized mortality rates above the upper 5% bound, and n is the number of forecasts including that for the starting point of the forecasts. Based on 5,000 simulation trials.

Figure 11
Bootstrapped *p*-Values of Realized Mortality Outcomes: Males Aged 65, 1980 Start, Horizon = 28 Years Ahead



Notes: Forecasts based on models estimated using English & Welsh male mortality data for ages 60–84 and years 1961–1980 assuming parameter uncertainty. Each figure shows the bootstrapped forecast CDF of the mortality rate in 2008, where the forecast is based on the density forecast made ''as if'' in 1980. Each black vertical line gives the realized mortality rate in 2008 and its associated *p*-value in terms of the forecast CDF. Based on 5,000 simulation trials.

5. MORTALITY PROBABILITY DENSITY FORECAST BACKTESTS: BACKTESTS BASED ON STATISTICAL HYPOTHESIS TESTS

The fourth and final class of backtests involves formal hypothesis tests based on comparisons of realized outcomes against forecasts of the relevant probability densities.[11]

To elaborate, suppose we wish to use data up to and including 1980 to evaluate the forecasted probability density function of the mortality rate of, say, 65-year-olds in 2008, involving a forecast horizon of 28 years ahead. A simple way to implement such a test is to use each model to forecast the cumulative density function (CDF) for the mortality rate in 2008, as of 1980, and compare how the realized mortality rate for 2008 compares with its forecasted distribution.

To carry out this backtest for any given model, we use the data from 1961 to 1980 to make a forecast of the CDF of the mortality rate of 65-year-olds in 2008.[12] The curves shown in Figure 11 are the CDFs for each of the six models. The null hypothesis is that the realized mortality rate for 65-year-old males in 2008 is consistent with the forecasted CDF. We then determine the p-value associated with the null hypothesis, which is obtained by taking the value of the CDF where the vertical line drawn up from the realized mortality rate on the x-axis crosses the cumulative density curve. Where the realized values fall into the left-hand tails, as all do in Figure 11, the reported p-values are those associated with one-sided tests of the null, which are more appropriate here given the evidence that longer-term forecasts are biased.

For the backtests illustrated in Figure 11, we get estimated p-values of 16.3%, 4.8%, 8.76%, 6.48%, 1.2%, and 7.72% for models M1, M2B, M3B, M5, M6, and M7, respectively. These results tell us that the null hypothesis that M1 generates adequate forecasts over a horizon of 28 years is associated with a probability of 16.3%, and so forth. In this case, all models ''pass'' at a 1% significance level.

Now suppose we are interested in carrying out a test of the models' 27-year-ahead mortality density forecasts. Two cases now can be considered: we can start in 1980 and carry out the test for a forecast of the 2007 mortality density, or we can start in 1981 and carry out the test for a forecast of the 2008 mortality density.[13] Along the same lines, if we wish to test the models' forecasts of 26-year-ahead density forecasts, we have three choices: start in 1980 and forecast the density for year 2006, start in 1981 and forecast the density for year 2007, and start in 1982 and forecast the density for year 2008.[14]

We can carry on in the same way for forecasts 23 years ahead, 22 years ahead, and so on, down to forecasts one year ahead. By the time we get to one-year-ahead forecasts, we would have 28 different possibilities (i.e., start in 1980, start in 1981, etc., up to start in 2007).

In all, this gives us $1 \times 28 + 2 \times 27 + 3 \times 26 + \cdots + 27 \times 2 + 28 \times 1 = 4{,}060$ possible such tests.

So for any given model, we can construct a sample of 28 different estimates of the p-value of the null associated with one-year-ahead forecasts, we can construct a sample of 27 different estimates of the p-value of the null associated with two-year-ahead forecasts, and so forth.

We can then construct plots of estimated p-values for any given horizon or set of horizons. Some examples are given in Figure 12, which shows plots of the p-values for 65-year-olds associated with forecast horizons of 10, 15, 20, and 25 years. The p-values are fairly variable, and notable differences exist across models: M7 performs fairly well, for example, while M6 performs poorly.

Figure 13 gives the corresponding results for 84-year-olds, and these forecasts perform much better with no ''problematic'' test results except in some cases for M1.

[11] Since we are testing forecasts of future mortality rates, we can also regard the tests considered here as similar to tests of q-forward forecasts using the different models; q-forwards are mortality rate forward contracts (Coughlan et al. 2007).

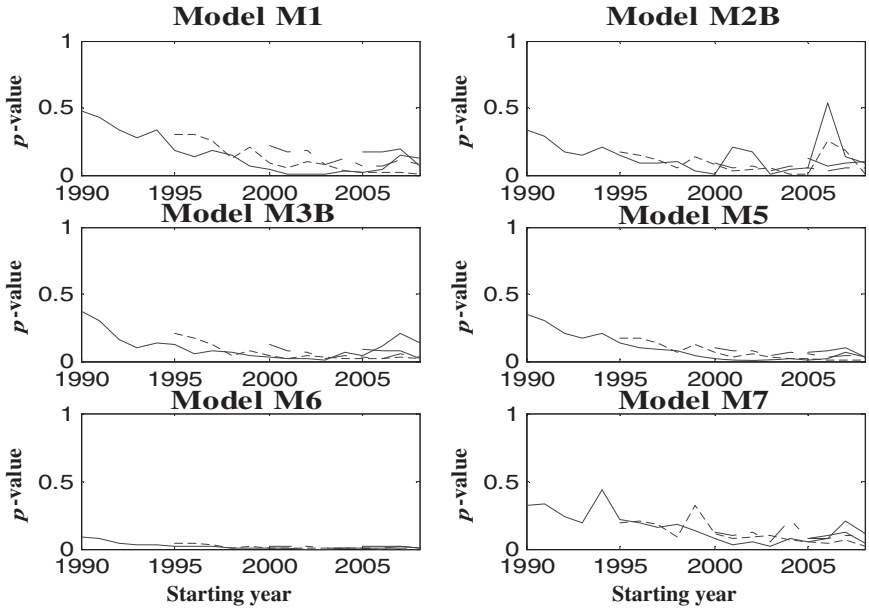
[12] Note that we are now dealing with PU forecasts only.

[13] In principle, both these tests should give much the same answer under the null hypothesis, and there is nothing to choose between them other than the somewhat extraneous consideration (for present purposes) that the latter test uses slightly later data.

[14] Again, in principle, each of these tests should yield similar results under the null.

Figure 12

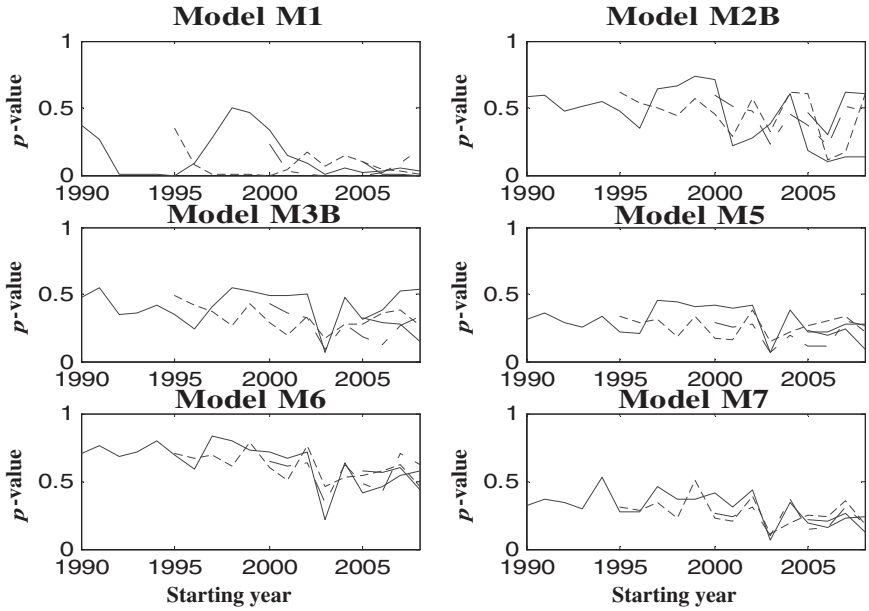
Various-Horizon *p*-Values of Realized Future Mortality Rates: Males Aged 65



Notes: Forecasts based on estimates using English & Welsh male mortality data for ages 60–84 and a rolling 20-year window assuming parameter uncertainty. The lines refer to *p*-values over horizons of h = 10, 15, 20, and 25 years, respectively, and the lengths of the lines correspond to the forecast horizons: the longest represents the *p*-values of the 10-year-horizon forecasts, etc. Based on 5,000 simulation trials.

Figure 13

Various-Horizon *p*-Values of Realized Future Mortality Rates: Males Aged 84



Notes: Forecasts based on estimates using English & Welsh male mortality data for ages 60–84 and a rolling 20-year window assuming parameter uncertainty. The lines refer to *p*-values over horizons of h = 10, 15, 20, and 25 years, respectively, and the lengths of the lines correspond to the forecast horizons: the longest represents the *p*-values of the 10-year-horizon forecasts, etc. Based on 5,000 simulation trials.

6. CONCLUSIONS

The purposes of this paper are (1) to set out a backtesting framework that can be used to evaluate the ex post forecasting performance of stochastic mortality models and (2) to illustrate this framework by evaluating the forecasting performance of a number of mortality models calibrated under a particular data set.

The backtesting framework presented here is based on the idea that forecast distributions should be compared against subsequently realized mortality outcomes: if the realized outcomes are compatible with their forecasted distributions, then this would suggest that the forecasts and the models that generated them are good ones; and if the forecast distributions and realized outcomes are incompatible, this would suggest that the forecasts and models are poor. We discussed four different classes of backtest building on this general idea: (1) backtests based on the convergence of forecasts through time toward the mortality rate(s) in a given year, (2) backtests based on the accuracy of forecasts over multiple horizons, (3) backtests based on the accuracy of forecasts over rolling fixed-length horizons, and (4) backtests based on formal hypothesis tests that involve comparisons of realized outcomes against forecasts of the relevant densities over specified horizons.

We generally find that the forecasts for age 65 show a clear bias that increases with the forecast horizon, whereas little such bias is found for the age-84 forecasts, and we often find more violations of the PC lower bounds than of their PU equivalents. By this criterion, the PU forecasts therefore often perform better.

As far as the individual models are concerned, we find that models M1, M3B, M5, M6, and M7 perform well most of the time, and (except for some problems with M1 and M6) there is relatively little to choose between these models.

Model M2B, however, repeatedly shows evidence of considerable instability. However, we should make clear that we have examined a particular version of M2 in this study and so cannot rule out the possibility that other specifications of, or extensions to, M2 might resolve the stability problem identified both here and elsewhere (e.g., Cairns et al. 2010; Dowd et al. 2010).

We also find that projections that incorporate parameter uncertainty are wider than those that do not.[15]

Finally, we would emphasize that these results are obtained using one particular data set—data for English & Welsh males—over limited sample periods. Accordingly, we make no claim for how these models might perform over other data sets or sample periods.

APPENDIX A: THE STOCHASTIC MORTALITY MODELS CONSIDERED IN THIS STUDY

Model M1

Model M1, the Lee-Carter model, postulates that the true underlying death rate, $m(t, x)$, satisfies

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

where the state variable $\kappa_t^{(2)}$ follows a one-dimensional random walk with drift

$$\kappa_t^{(2)} = \kappa_{t-1}^{(2)} + \mu + CZ_t^{(2)} \tag{A.2}$$

in which μ is a constant drift term, C a constant volatility, and $Z_t^{(2)}$ a one-dimensional iid N(0,1) error. This model also satisfies the identifiability constraints:

$$\sum_t \kappa_t^{(2)} = 0 \text{ and } \sum_x \beta_x^{(2)} = 1. \tag{A.3}$$

[15] An interesting extension is suggested by Li et al. (2009). They replace the Poisson treatment of the deaths counts, $D(t, x)$, with a negative binomial distribution that allows for heterogeneity in population characteristics within each (t, x) data cell. Using M1 applied to U.S. and Canadian data, they find that this leads to wider prediction intervals than one obtains with the Poisson. We might expect, therefore, that replacing the Poisson with a negative binomial would have a similar effect on the other models considered here.

Model M2B

Model M2B is an extension of Lee-Carter that allows for a cohort effect. It postulates that $m(t, x)$ satisfies

$$\log m(t, x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_c^{(3)}, \tag{A.4}$$

where the state variable $\kappa_t^{(2)}$ follows (A.2) and $\gamma_c^{(3)}$ is a cohort effect where $c = t - x$ is the year of birth. We follow Cairns et al. (2008) and CMI (2007) and model $\gamma_c^{(3)}$ as an ARIMA(1,1,0) process:

$$\Delta\gamma_c^{(3)} = \mu^{(\gamma)} + \alpha^{(\gamma)}(\Delta\gamma_{c-1}^{(3)} - \mu^{(\gamma)}) + \sigma^{(\gamma)}Z_c^{(\gamma)} \tag{A.5}$$

with parameters $\mu^{(\gamma)}$, $\alpha^{(\gamma)}$, and $\sigma^{(\gamma)}$, and $Z_c^{(\gamma)}$ is iid N(0,1). This model satisfies the identifiability constraints:

$$\sum_t \kappa_t^{(2)} = 0, \sum_x \beta_x^{(2)} = 1, \sum_c \gamma_c^{(3)} = 0, \text{ and } \sum_x \beta_x^{(3)} = 1. \tag{A.6}$$

Model M3B

This model is a simplified version of M2B that postulates

$$\log m(t, x) = \beta_x^{(1)} + \frac{1}{n_a}\kappa_t^{(2)} + \frac{1}{n_a}\gamma_c^{(3)}, \tag{A.7}$$

where n_a is the number of ages on which the parameters are estimated, and the variables (including the cohort effect) are the same as for M2B. This model satisfies the constraints

$$\sum_t \kappa_t^{(2)} = 0 \text{ and } \sum_{x,t} \gamma_{t-x}^{(3)} = 0. \tag{A.8}$$

Given these constraints, we let $\bar{\beta}_x^{(1)} = 20^{-1}\,\Sigma_t \log m(t, x)$—bearing in mind that our lookback sample window has a length of 20 years—and obtain

$$\delta = -\frac{\sum_x (x - \bar{x})(\beta_x^{(1)} - \bar{\beta}_x^{(1)})}{\sum_x (x - \bar{x})^2} \tag{A.9}$$

and thence obtain the following revised parameter estimates:

$$\begin{aligned}
\tilde{\kappa}_t^{(2)} &= \kappa_t^{(2)} - \delta(t - \bar{t}),\\
\tilde{\gamma}_{t-x}^{(3)} &= \gamma_{t-x}^{(3)} + \delta((t - \bar{t}) - (x - \bar{x})),\\
\tilde{\beta}_x^{(1)} &= \beta_x^{(1)} + \delta(x - \bar{x}),
\end{aligned} \tag{A.10}$$

which are the ones on which the forecasts are based.

Model M5

Model M5 is a reparameterized version of the CBD two-factor mortality model (Cairns et al. 2006) and postulates that the mortality rate $q(t, x)$ satisfies

$$\text{logit } q(t, x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}), \tag{A.11}$$

where $q(t, x) = 1 - \exp(-m(t, x))$, $\bar{x}$ is the average of the ages used in the dataset, and the state variables follow a two-dimensional random walk with drift:

$$\kappa_t = \kappa_{t-1} + \mu + CZ_t \tag{A.12}$$

in which μ is a constant 2×1 drift vector, C is a constant 2×2 upper triangular "volatility" matrix (to be precise, the Choleski "square root" matrix of the variance-covariance matrix), and Z_t is a two-dimensional standard normal variable, each component of which is independent of the other.

Model M6

M6 is a generalized version of M5 with a cohort effect:

$$\text{logit } q(t, x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_c^{(3)}, \tag{A.13}$$

where the κ_t process follows (A.12), and the $\gamma_c^{(3)}$ process follows (A.5). This model satisfies the identifiability constraints:

$$\sum_{c \in C} \gamma_c^{(3)} = 0 \text{ and } \sum_{c \in C} c\gamma_c^{(3)} = 0. \tag{A.14}$$

Model M7

M7 is a second generalized version of M5 with a cohort effect:

$$\text{logit } q(t, x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}((x - \bar{x})^2 - \sigma_x^2) + \gamma_c^{(4)}, \tag{A.15}$$

where the state variables κ_t follow a three-dimensional random walk with drift, σ_x^2 is the variance of the age range used in the dataset, and $\gamma_c^{(4)}$ is a cohort effect that is modeled as an AR(1) process. This model satisfies the identifiability constraints:

$$\sum_{c \in C} \gamma_c^{(4)} = 0, \sum_{c \in C} c\gamma_c^{(4)} = 0, \text{ and } \sum_{c \in C} c^2\gamma_c^{(4)} = 0. \tag{A.16}$$

More details on the models can be found in Cairns et al. (2009).

APPENDIX B: SIMULATING MODEL PARAMETERS UNDER CONDITIONS OF PARAMETER UNCERTAINTY

This appendix explains the procedure used to simulate the values of the parameters driving the κ_t process and, where the model involves a cohort effect γ_c, the parameters of the process that drives γ_c. The approach used is a Bayesian one based on noninformative Jeffreys prior distributions (Cairns 2000; Cairns et al. 2006).

Simulating the Values of the Parameters Driving the κ_t Process

Recall from Appendix A that each of the models has an underlying random walk with drift

$$\kappa_t = \kappa_{t-1} + \mu + CZ_t, \tag{B.1}$$

where, in the case of M1, M2B, and M3B, κ_t, μ, and V are scalars; in the case of M5 and M6, κ_t and μ are 2×1 vectors and a V is 2×2 matrix; and in the case of M7, κ_t and μ are 3×1 vectors, and V is a 3×3 matrix. For the sake of generality, we refer to μ as a vector and V as a matrix.

The parameters of (B.1) are estimated by MLE using a sample of size $n = 20$ for the results presented in this paper.

We first simulate V from its posterior distribution, the Wishart $(n - 1, n^{-1}\hat{V}^{-1})$ distribution. To do so, we carry out the following steps:

- We simulate $n - 1$ i.i.d. vectors $\alpha_1, \ldots, \alpha_{n-1}$ from a multivariate normal distribution with mean vector 0 and covariance matrix $n^{-1}\hat{V}^{-1}$. These vectors will have dimension 1×1 for M1, M2B, and M3B, dimension 2×1 for M5 and M6, and dimension 3×1 for M7.
- We construct the matrix $X = \Sigma_{i=1}^{n-1} \alpha_i \alpha_i^T$.
- We then invert X to obtain our simulated positive-definite covariance matrix $V(= X^{-1})$.

Having obtained our simulated V matrix, we simulate the μ vector from $MVN(\hat{\mu}, n^{-1}V)$.

Simulating the Values of the Parameters Driving the γ_c Process

Recall that c is the year of birth. We first note that, for model M7, the cohort effect $\gamma_c^{(4)}$ follows an AR(1) process, and for models M2B, M3B, and M5, the cohort effect $\gamma_c^{(3)}$ follows an ARIMA(1,1,0) process. This latter process implies that the first difference of $\gamma_c^{(3)}$, $\Delta\gamma_c^{(3)}$, follows an AR(1). Hence our task is to simulate the values of the parameters $\mu^{(\gamma)}$, $\alpha^{(\gamma)}$, and $\sigma^{(\gamma)}$ in the following equation:

$$\gamma_c = \mu^{(\gamma)} + \alpha^{(\gamma)}(\gamma_{c-1} - \mu^{(\gamma)}) + \sigma^{(\gamma)}\varepsilon_c, \tag{B.2}$$

and γ_c is, as appropriate, either $\Delta\gamma_c^{(3)}$ or $\gamma_c^{(4)}$. The method we use is taken from Cairns (2000, pp. 320–321) and is based on a Jeffreys prior distribution for $(\mu^{(\gamma)}, \alpha^{(\gamma)}, \sigma^{(\gamma)})$ and MLE estimates of these parameters $(\hat{\mu}^{(\gamma)}, \hat{\alpha}^{(\gamma)}, \hat{\sigma}^{(\gamma)})$. This involves the following steps:

- We simulate the value of $\alpha^{(\gamma)}$ from $\hat{\alpha}^{(\gamma)}$ and n from its posterior distribution $f(\alpha^{(\gamma)}) = ((\alpha^{(\gamma)})^2 - 2\alpha^{(\gamma)}\hat{\alpha}^{(\gamma)} + 1)^{-(n-1)/2}$ using a suitable algorithm (e.g., the rejection method).
- We then simulate $X \sim \chi^2_{n-1}$ and obtain a simulated value of $(\sigma^{(\gamma)})^2$ from $(\sigma^{(\gamma)})^2 = (n-1)(\hat{\sigma}^{(\gamma)})^2(1 + (\alpha^{(\gamma)} - \hat{\alpha}^{(\gamma)})^2/(1 - (\hat{\alpha}^{(\gamma)})^2))/X$.
- Finally, we simulate $Z \sim N(0, 1)$ and obtain a simulated value of $\mu^{(\gamma)}$ from $\mu^{(\gamma)} = \hat{\mu}^{(\gamma)} + \sqrt{(\sigma^{(\gamma)})^2/(n-1)/(1-\alpha^{(\gamma)})^2} \times Z$.

Further details are provided in Cairns (2000, pp. 320–321).

7. ACKNOWLEDGMENTS

The authors thank Lixia Loh, Sharif Mozumder, and Liang Zhao for excellent research assistance, and an anonymous referee for helpful comments. The usual caveat applies. Additional information is available upon request. This report has been partially prepared by the Pension Advisory Group, and not by any research department, of JPMorgan Chase & Co. and its subsidiaries (''JPMorgan''). Information herein is obtained from sources believed to be reliable, but JPMorgan does not warrant its completeness or accuracy. Opinions and estimates constitute JPMorgan's judgment and are subject to change without notice. Past performance is not indicative of future results. This material is provided for informational purposes only and is not intended as a recommendation or an offer or solicitation for the purchase or sale of any security or financial instrument.

REFERENCES

BLAKE, D., K. DOWD, AND A. J. G. CAIRNS. 2008. Longevity Risk and the Grim Reaper's Toxic Tail: The Survivor Fan Charts. *Insurance: Mathematics and Economics* 42: 1062–1068.

CAIRNS, A. J. G. 2000. A Discussion of Parameter and Model Uncertainty in Insurance. *Insurance: Mathematics and Economics* 27: 313–330.

CAIRNS, A. J. G., D. BLAKE, AND K. DOWD. 2006. 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., D. BLAKE, K. DOWD, G. D. COUGHLAN, D. EPSTEIN, AND M. KHALAF-ALLAH. 2010. A Framework for Forecasting Mortality Rates with an Application to Six Stochastic Mortality Models. Pensions Institute Discussion Paper PI-0801.

CAIRNS, A. J. G., D. BLAKE, K. DOWD, G. D. COUGHLAN, D. EPSTEIN, A. ONG, AND I. BALEVICH. 2009. A Quantitative Comparison of Stochastic Mortality Models Using Data from England & Wales and the United States. *North American Actuarial Journal* 13(1): 1–35.

CONTINUOUS MORTALITY INVESTIGATION. 2006. Stochastic Projection Methodologies: Further Progress and P-Spline Model Features, Example Results and Implications. Working Paper 20.

CONTINUOUS MORTALITY INVESTIGATION. 2007. Stochastic Projection Methodologies: Lee-Carter Model Features, Example Results and Implications. Working Paper 25.

COUGHLAN, G., D. EPSTEIN, A. SINHA, AND P. HONIG. 2007. *q*-Forwards: Derivatives for Transferring Longevity and Mortality Risks. Available at www.jpmorgan.com/lifemetrics.

CURRIE, I. D., M. DURBAN, AND P. H. C. EILERS. 2004. Smoothing and Forecasting Mortality Rates. *Statistical Modelling* 4: 279–298.

DOWD, K., A. J. G. CAIRNS, AND D. BLAKE. 2006. Mortality-Dependent Measures of Financial Risk. *Insurance: Mathematics and Economics* 38: 427–440.

DOWD, K., CAIRNS, A. J. G., D. BLAKE, G. D. COUGHLAN, D. EPSTEIN, AND M. KHALAF-ALLAH. 2010. Evaluating the Goodness of Fit of Stochastic Mortality Models. Pensions Institute Discussion Paper PI-0802. Forthcoming in *Insurance: Mathematics and Economics*.

EMBRECHTS, P., C. KLÜPPELBERG, AND T. MIKOSCH. 1997. *Modelling Extreme Events for Insurance and Finance*. Berlin: Springer.

JACOBSEN, R. N., N. KEIDING, AND E. LYNGE. 2002. Long-Term Mortality Trends behind Low Life Expectancy of Danish Women. *Journal of Epidemiology and Community Health* 56: 205–208.

KEILMAN, N. 1997. Ex Post Errors in Official Population Forecasts in Industrialized Countries. *Journal of Official Statistics (Statistics Sweden)* 13: 245–247.

KEILMAN, N. 1998. How Accurate Are the United Nations World Population Projections? *Population and Development Review* 24: 15–41.

KOISSI, M.-C., A. SHAPIRO, AND G. HÖHNÄS. 2006. Evaluating and Extending the Lee-Carter Model for Mortality Forecasting: Bootstrap Confidence Interval. *Insurance: Mathematics and Economics* 38: 1–20.

LEE, R. D., AND L. R. CARTER. 1992. Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association* 87: 659–675.

LEE, R. D., AND T. MILLER. 2001. Evaluating the Performance of the Lee-Carter Method for Forecasting Mortality. *Demography* 38: 537–549.

LI, J. S.-H., M. R. HARDY, AND K. S. TAN. 2009. Uncertainty in Mortality Forecasting: An Extension of the Classical Lee-Carter Approach. *ASTIN Bulletin* 39: 137–164.

LONGSTAFF, F. A., AND E. S. SCHWARTZ. 1992. Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model. *Journal of Finance* 47: 1259–1282.

NATIONAL RESEARCH COUNCIL. 2000. *Beyond Six Billion: Forecasting the World's Population,* edited by J. Bongaerts and R. A. Bulatao. Washington, DC: National Academy Press.

OSMOND, C. 1985. Using Age, Period and Cohort Models to Estimate Future Mortality Rates. *International Journal of Epidemiology* 14: 124–29.

RENSHAW, A. E., AND S. HABERMAN. 2006. A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. *Insurance: Mathematics and Economics* 38: 556–570.

VASICEK, O. 1977. An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics* 5: 177–188.

WILMOTH, J. R. 1998. Is the Pace of Japanese Mortality Decline Converging toward International Trends? *Population and Development Review* 24: 593–600.

Discussions on this paper can be submitted until January 1, 2011. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.