



UvA-DARE (Digital Academic Repository)

Accounting for COVID-19-Type shocks in mortality modeling

A comparative study

Schnürch, S.; Kleinow, T.; Wagner, A.

DOI

[10.1017/dem.2023.9](https://doi.org/10.1017/dem.2023.9)

Publication date

2023

Document Version

Final published version

Published in

Journal of Demographic Economics

License

Article 25fa Dutch Copyright Act (<https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment>)

[Link to publication](#)

Citation for published version (APA):

Schnürch, S., Kleinow, T., & Wagner, A. (2023). Accounting for COVID-19-Type shocks in mortality modeling: A comparative study. *Journal of Demographic Economics*, 89(3), 483-512. <https://doi.org/10.1017/dem.2023.9>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

RESEARCH PAPER

Accounting for COVID-19-type shocks in mortality modeling: a comparative study

Simon Schnürch^{1,2}, Torsten Kleinow³ and Andreas Wagner^{1,4}

¹Department of Financial Mathematics, Fraunhofer Institute for Industrial Mathematics ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany, ²Department of Mathematics, University of Kaiserslautern, Gottlieb-Daimler-Straße 48, 67663 Kaiserslautern, Germany, ³Department of Actuarial Mathematics and Statistics and the Maxwell Institute for Mathematical Sciences, School of Mathematical and Computer Sciences, Heriot-Watt University, EH14 4AS Edinburgh, UK and ⁴Faculty of Management Science and Engineering, Karlsruhe University of Applied Sciences, Moltkestraße 30, 76133 Karlsruhe, Germany

Corresponding author: E-mail: schnuerch.itwm@web.de

(Received 6 April 2023; revised 6 April 2023; accepted 6 April 2023)

Abstract

Mortality shocks such as the one induced by the COVID-19 pandemic have substantial impact on mortality models. We describe how to deal with them in the period effect of the Lee–Carter model. The main idea is to not rely on the usual normal distribution assumption as it is not always justified. We consider a mixture distribution model based on the peaks-over-threshold method, a jump model, and a regime switching model and introduce a modified calibration procedure to account for the fact that varying amounts of data are necessary for calibrating different parts of these models. We perform an extensive empirical study for nine European countries, comparing the models with respect to their parameters, quality of fit, and forecasting performance. Moreover, we define five exemplary scenarios regarding the future development of pandemic-related mortality. As a result of our evaluations, we recommend the peaks-over-threshold approach for applications with a possibility of extreme mortality events.

Keywords: COVID-19; Lee–Carter model; mortality forecasting; mortality modeling; mortality shocks

JEL classification: J11; C53; G22

1. Introduction

Globalization, population growth, and other factors have increased the likelihood for the occurrence of epidemics and pandemics [Engel and Ziegler (2020)], with COVID-19 being the most recent example. A severe pandemic is just one of several low-probability, high-impact events which may lead to a sudden increase in mortality rates, a so-called mortality shock. Mortality shocks lead to changes in mortality data, which in turn can influence model parameters and forecasts [Schnürch *et al.* (2022)]. This is relevant for many applications, including the pricing of mortality catastrophe bonds as well as risk management and reserving decisions in life insurance companies.

A central question with respect to mortality shocks is if and how to treat them in mortality models, regarding both their appearance in past data to which models are calibrated and the possibility of their appearance in future observations. We especially focus on the popular stochastic mortality model by Lee and Carter (1992, LC). Forecasts under this model are usually obtained by a random walk with drift, which corresponds to a normal distribution assumption. As we demonstrate, mortality shocks can lead to violations of this assumption with potential consequences for forecasts and the quantification of their uncertainty.

One way of addressing this is to suitably adjust the data by outlier adjustment methods. However, depending on the application, it might be questionable to remove genuine observations of extreme mortality from the data. Under the Solvency II framework, for example, solvency capital requirements are calculated as a value-at-risk, for which a realistic evaluation of tail behavior is vital. Moreover, since the COVID-19 pandemic is by now expected to influence mortality data of at least two years (2020 and 2021), exclusion of extreme data would lead to quite a lot of recent information not entering model calibration.

Therefore, it could be more reasonable to not (only) change the data but to change the model by deviating from the normal distribution assumption. This can be done in a generic way, switching to another distribution family with heavier upper tail. Here, we propose and investigate the lognormal distribution. Alternatively, it is possible to explicitly model the occurrence of extreme mortality events, for which we present three approaches (“shock models”) from the literature: a mixture distribution based on the peaks-over-threshold method, a discrete jump diffusion and a regime switching model. They have different theoretical properties and all of them can be useful for describing historical data, but empirically we find the mixture distribution to be most appropriate with respect to several criteria.

Our main contribution to the literature is twofold:

- We propose an adjusted calibration procedure for the shock models, using long-term data to fit the shock parameters and short-term data to fit the remaining parameters, and demonstrate that it clearly improves backtest performance.
- We conduct an extensive empirical evaluation of the different modeling approaches with respect to quality of fit, forecasting performance, and differences in predicted period effects as well as term assurance and annuity values. We introduce five exemplary scenarios for the further development of pandemic-related mortality and show how the models would perform under these scenarios.

Our empirical evaluations take place on an updated version of the data set used by Schnürch *et al.* (2022) and we refer to this paper for a more detailed description of the data and notation. We get yearly death rates $m_{x,t}^i$ directly from the Human Mortality Database (2021). For years in which these are unavailable yet, we calculate them from weekly death counts obtained from Short Term Mortality Fluctuations (2021). Here, we denote age by $x \in \mathcal{X} = \{x_1, \dots, x_A\}$ and consider 5-year age groups so that $x_1 = 35 - 39$, $x_2 = 40 - 44$, ..., $x_A = 90 +$. The population is denoted by $i \in \mathcal{P} := \{1, 2, \dots, P\}$ with $P = 27$ as we consider the male, female, and total populations of 9 European countries.¹ The availability of years depends on the

¹To save space, we only show results for total populations in some evaluations.

country, which we usually ignore in our notation and simply denote the calendar year by $t \in \mathcal{T} := \{t_1, \dots, t_Y\}$.

We proceed as follows: section 2 introduces the LC model along with different ways of forecasting its period effect. For the approaches explicitly modeling the possibility of mortality shocks, we describe a new, adjusted calibration procedure. In section 3 we present empirical results on the normal distribution assumption for the period effect increments and evaluations of the models with respect to quality of fit, forecasts, and forecasting performance in a backtest and under different future scenarios. Section 4 concludes.

2. Methodology

2.1. The Lee-Carter model

Lee and Carter (1992) model logarithmic death rates for a life aged x in year t belonging to population i as

$$\log m_{x,t}^i = \alpha_x^i + \beta_x^i \kappa_t^i + \varepsilon_{x,t}^i, \tag{1}$$

with age-specific base mortality level α_x^i , period effect κ_t^i related to the general development over time, and age effect β_x^i describing the impact of this development on different ages. The error terms $\varepsilon_{x,t}^i$ are assumed to be independent, homoskedastic, and normally distributed. In the following, we suppress the dependence on the population i in our notation.

The model (1) is calibrated via singular value decomposition. The constraints

$$\sum_{x=x_1}^{x_A} \beta_x = 1, \kappa_{t_1} = 0 \tag{2}$$

are imposed to make its parameters identifiable. Forecasts of future mortality rates are obtained by choosing a suitable model for the estimated period effects $(\hat{\kappa}_t)_{t \in \mathcal{T}}$, usually for their first differences $\Delta \hat{\kappa}_t := \hat{\kappa}_t - \hat{\kappa}_{t-1}$.

The choice of such a time series model depends on its compatibility with observed data but also on the expectations and preferences of the modeler. In the following, we present several possibilities for this choice. An overview is given in Table 1.

2.2. Forecasting the period effect with an ARIMA model

The simplest and most often used approach is to model $(\hat{\kappa}_t)_{t \in \mathcal{T}}$ as a random walk with drift (RWD), i.e.,

$$\Delta \hat{\kappa}_t = e_t \text{ with } e_t \sim \mathcal{N}(\mu, \sigma^2) \text{ i.i.d.} \tag{3}$$

This yields explicit formulae for both point and interval forecasts of the period effect, which are then inserted into (1) to obtain death rate forecasts. We refer to Schnürch *et al.* (2022), who provide details on the procedure and demonstrate that a mortality shock in the forecast jump-off year influences RWD point forecasts and a mortality shock at any time in the calibration data set influences interval forecasts.

Table 1. Overview of the different modeling approaches for $\hat{\kappa}_t$

Model name (abbreviation)	Short description	Treatment of shocks
rwd	Random walk with drift	Does not ignore shocks but also does not explicitly allow for them
auto.arima	General ARIMA model	Does not ignore shocks but also does not explicitly allow for them
best_estimate	Replace values in shock year by best estimate	Completely ignores shock in all model parameters (including α_x and β_x)
intervention	Introduce dummy for shock year	Shock year effect is captured by a separate parameter
lognormal	Assume lognormal distribution of period effect increments	Implicit modeling: lognormal distribution has heavier upper tail than the usual normal distribution
pot	Peaks-over-threshold model	Shocks are explicitly modeled using a generalized Pareto distribution
jump	Model with transitory, normally distributed jumps	Shocks are explicitly modeled as jumps with Bernoulli occurrence and normal severity
rs	Binary regime switching model	Shocks are explicitly modeled as regime changes

Although the RWD has been demonstrated to work quite well in the literature [Lee and Carter (1992); Tuljapurkar *et al.* (2000)], it has a few drawbacks, for example, its strong dependence on $\hat{\kappa}_{t_y}$, the period effect in the jump-off year. A natural idea is to replace the RWD by a general autoregressive integrated moving average (ARIMA) model, which potentially places less weight on the last period effect and also yields a better fit. ARIMA models have three hyperparameters, the autoregressive order p , the order of integration d , and the moving average order q , which we choose by the auto.arima algorithm: for each population, we consider all ARIMA(p, d, q) process specifications with $p \leq 5$, $d \leq 1$ and $q \leq 5$ and select the one which strikes the best balance between low calibration error and parsimony. We refer to Hyndman and Khandakar (2008, Section 3) for a detailed description of the algorithm as well as the R [R Core Team (2019)] implementation we use. Similarly as for an RWD, a mortality shock in the calibration data set of an ARIMA(p, d, q) model will persist in its forecasts if $p > 0$ or $d > 0$, which is the case for all ARIMA models the auto.arima algorithm has returned in our numerical experiments.

2.3. Dealing with mortality shocks in an ARIMA setting

In this section, we describe six alternative methods to explicitly deal with outliers in the period effect.

One way of addressing the occurrence of mortality shocks is to reduce or completely remove their effect on calibration, so that the model is trained with a focus on “normal” data. Therefore, it should only be used to make forecasts conditional on the assumption

that no major mortality shocks will occur in the future or if the possibility of such shocks is not relevant for the considered application.

Replacing by a best estimate. The most radical way of reducing the effect of mortality shocks on model calibration is to exclude years in which they occur from the data, which has a similar effect as replacing the observations in these years by best estimates based on shock-free data. For the example of COVID-19, we calibrate an LC model on data up to 2019, obtain forecasts for 2020 by an RWD and recalibrate the model on the data including these 2020 forecasts. This approach has been investigated and compared to the RWD without data exclusion by Schnürch *et al.* (2022). It obviously prevents the excluded mortality shock from influencing mortality forecasts in any way.

Intervention model. Alternatively, one can apply an intervention model to the period effect time series, as Lee and Carter (1992) have done to diminish the influence of the 1918 pandemic. This means that a binary dummy is added to the RWD model (3), yielding

$$\Delta \hat{\kappa}_t = \varphi 1_{t=t^{\text{out}}} + e_t, \tag{4}$$

where $t^{\text{out}} := \operatorname{argmax}_{t=t_2, \dots, t_Y} \Delta \hat{\kappa}_t$ is the year with the largest (upward) mortality shock and the maximum likelihood estimate for φ is $\Delta \hat{\kappa}_{t^{\text{out}}} - \hat{\mu}$. For the calibration period $t \in \{1991, \dots, 2020\}$, we usually get $t^{\text{out}} = 2020$ so that the parameter φ absorbs the mortality shock induced by the COVID-19 pandemic. Therefore, this intervention model can be expected to behave very similarly to our best estimate approach which directly ignores 2020. In particular, it only allows the mortality shock to influence forecasts via the α_x and β_x parameters, which are calibrated on the original data including the shock. However, changes in these parameters in response to a COVID-19-type shock are expected to be small [Schnürch *et al.* (2022), Proposition 1].

More sophisticated methods are available, including more general intervention models [Box and Tiao (1975)] or techniques from time series outlier detection and adjustment [Li and Chan (2005, 2007)]. However, we restrict our attention to the best estimate replacement and the simple intervention model (4) as we deem these sufficient to illustrate the consequences of excluding mortality shocks from the model calibration.

Another way of dealing with mortality shocks is to include them in the calibration data and adapt the model accordingly. More precisely, we adjust the distribution assumption.

Lognormal distribution. A natural idea is to simply replace the normal distribution $\mathcal{N}(0, \sigma^2)$ in (3) by some asymmetric, unimodal distribution with heavier upper tail. There are several distributions with this property in the literature. We choose the lognormal distribution, which is known to be suitable for a lot of applications in finance such as modeling stock prices or interest rates.

As the lognormal distribution is only supported on the positive real line and $\Delta \hat{\kappa}_t$ is usually negative for many t , some data transformation is necessary. We apply

$$\Delta \hat{\kappa}_t \mapsto \Delta \hat{\kappa}_t + \left| \min_{\tilde{t}=t_2, \dots, t_Y} \Delta \hat{\kappa}_{\tilde{t}} \right| + \lambda \text{ for } t = t_2, \dots, t_Y \tag{5}$$

with $\lambda > 0$. The entire approach is equivalent to a Box–Cox transformation [Box and Cox (1964)] with non-zero shift parameter $\left| \min_{i=t_2, \dots, t_Y} \Delta \hat{\kappa}_i \right| + \lambda$.

We calibrate λ along with the two parameters of the lognormal distribution by maximum likelihood estimation. With the calibrated model, we perform Monte Carlo simulation (with 10^4 paths) and apply the inverse of (5) to obtain period effect increment forecasts, which are then converted into death rate forecasts. As the forecast for $\hat{\kappa}_{t_Y+1}$ is based on the unmodified value of $\hat{\kappa}_{t_Y}$, a shock to the period effect in the jump-off year will persist in forecasts.

Of course, it is also possible to change the distribution assumption so as to *explicitly* account for mortality shocks. We consider three methods from the literature, which are all based on the idea that there are two different underlying states governing mortality, a normal state and a shock state. This idea is implemented more or less directly.

Peaks-over-threshold. Chen and Cummins (2010) propose to apply the peaks-over-threshold (POT) method from extreme value theory to the first differenced period effect $\Delta \hat{\kappa}_t$, assuming that it follows a two-component mixture distribution. More precisely, a threshold $u \in \mathbb{R}$ is specified below which $\Delta \hat{\kappa}_t$ is modeled by a normal distribution, whereas it is modeled by a generalized Pareto distribution (GPD) above this threshold. We denote this by

$$\begin{aligned} \Delta \hat{\kappa}_t \mid \Delta \hat{\kappa}_t < u &\sim \mathcal{N}(\mu, \sigma^2), \\ \Delta \hat{\kappa}_t \mid \Delta \hat{\kappa}_t \geq u &\sim G(\xi, \zeta), \end{aligned} \tag{6}$$

where $G(\xi, \zeta)$ is a GPD with shape parameter $\xi \in \mathbb{R}$ and scaling parameter $\zeta > 0$. The period effect follows an RWD during normal times, but the model allows for the fact that shocks can occur and lead to non-normal behavior by approximating the upper tail of the distribution with a GPD. This is justified by the Pickands–Balkema–de Haan theorem from extreme value theory [Balkema and de Haan (1974); Pickands (1975)]. Similar mixture models have been successfully applied both on synthetic [Mendes and Lopes (2004)] and on real data [stock index returns, Behrens *et al.* (2004)].

A critical aspect of this method is the choice of the threshold u , which should not be too large so that the GPD parameters are calibrated on sufficiently many observations, but it should also not be too small so that only a minor share of the observations falls in the non-normal part of the distribution. Many procedures to estimate the threshold have been proposed in the literature [Scarrott and MacDonald (2012)]. Here, we follow Chen and Cummins (2010); Mendes and Lopes (2004) and calculate the profile likelihood

$$\mathcal{L}_p(u) = \sup_{\mu, \sigma, \xi, \zeta} \mathcal{L}(\mu, \sigma, \xi, \zeta; u), \tag{7}$$

where \mathcal{L} denotes the likelihood function of model (6). We try all values $u \in \{q_{85}(\Delta \hat{\kappa}_t), q_{86}(\Delta \hat{\kappa}_t), \dots, q_{95}(\Delta \hat{\kappa}_t)\}$, where $q_z(\Delta \hat{\kappa}_t)$ denotes the empirical $z\%$ quantile of $\Delta \hat{\kappa}_t$, and we choose u to maximize $\mathcal{L}_p(u)$. Then, we calibrate the parameters μ , σ , ξ , and ζ conditionally on the fixed u by penalized maximum likelihood estimation. The likelihood is penalized as proposed by Coles and Dixon

(1999) in order to stabilize the calibration.² We refer to Chen and Cummins (2010) and section 2.4 below for further details on the calibration algorithm. Chen and Cummins (2010) also describe how to simulate from their model, which we use to obtain point and interval death rate forecasts by Monte Carlo simulation (with 10^6 paths). In particular, analogously to the lognormal distribution approach, forecasts for $\hat{\kappa}_{t_Y+1}$ are based on the unmodified value of $\hat{\kappa}_{t_Y}$ and a shock to the period effect in the jump-off year t_Y will persist in forecasts.

Jump model. Another possibility to account for shocks is to allow for transitory mortality jumps. Several applications of jumps have been investigated in the literature both on discrete-time and continuous-time mortality models, see Regis and Jevtić (2022) for a recent overview, which also includes multi-population models. Here, we follow Chen and Cox (2009), who propose

$$\Delta\hat{\kappa}_t = e_t + W_t N_t - W_{t-1} N_{t-1}, \tag{8}$$

where $e_t \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d., the jump severities W_t follow i.i.d. $\mathcal{N}(m, s^2)$ distributions and N_t are i.i.d. Bernoulli random variables with jump probability $p \in (0, 1)$, i.e., $N_t = 1$ means there is a jump in mortality at time t and $N_t = 0$ means there is none. Independence is assumed between e_t , W_t , and N_t . It is easy to see that $\hat{\kappa}_t$ follows an RWD as long as there is no jump and that it returns to this RWD after a jump has taken place, unless $N_t = N_{t-1} = 1$. This latter case corresponds to two consecutive jumps or one longer-lasting jump and by (8) leads to an increase in volatility of $\Delta\hat{\kappa}_t$ from $\sqrt{\sigma^2 + s^2}$ to $\sqrt{\sigma^2 + 2s^2}$.

Although it is possible for longer-lasting shocks to occur under the jump model, its conceptual idea is to model transitory shock events which only influence mortality for one time period (one year). In particular, we refrain from modeling a serial correlation of N_t , i.e., a non-trivial dependence structure of shock occurrence over time.

In contrast to the POT model, we do not only consider the possibility of upward jumps here, as we allow $m \leq 0$. However, we expect that $m > 0$ for most populations because transitory mortality jumps usually move in an upward direction, whereas mortality improvements often play out over a longer time period.

The model is calibrated by conditional maximum likelihood estimation as described by Chen and Cox (2009). Point and interval forecasts are obtained via Monte Carlo simulation (with $J = 10^6$ paths) from (8). Estimates for the realized values of the unobservable W_{t_Y} and N_{t_Y} must be available in order to obtain the first forecast $\hat{\kappa}_{t_Y+1}$. We propose the following approach to calculate these estimates: Set $N_{t_Y}^a := 1$ and

$$W_{t_Y}^a := \mathbb{E}(W_{t_Y} | \Delta\hat{\kappa}_{t_Y}) \tag{9}$$

on Monte Carlo paths $a = 1, \dots, \pi J$, where

$$\pi := \mathbb{P}(N_{t_Y} = 1 | \Delta\hat{\kappa}_{t_Y}) \tag{10}$$

and πJ is rounded to the nearest integer. On the remaining Monte Carlo paths, $a = \pi J + 1, \dots, J$, set $N_{t_Y}^a := 0$ and the value of $W_{t_Y}^a$ does not matter. Applying Bayes' theorem to

²We use the R [R Core Team (2019)] implementation provided in the POT package [Ribatet and Dutang (2019)] for maximizing the penalized likelihood.

(10), we obtain

$$\begin{aligned} \pi &= \frac{f(\Delta\hat{\kappa}_{t_Y} | N_{t_Y} = 1) \cdot \mathbb{P}(N_{t_Y} = 1)}{f(\Delta\hat{\kappa}_{t_Y})} \\ &= \frac{((1 - p) \cdot \phi_{\mu+m, \sigma^2+s^2}(\Delta\hat{\kappa}_{t_Y}) + p \cdot \phi_{\mu, \sigma^2+2s^2}(\Delta\hat{\kappa}_{t_Y})) \cdot p}{f(\hat{\kappa}_{t_Y})}, \end{aligned} \tag{11}$$

where f denotes the (conditional) density function of $\Delta\hat{\kappa}_{t_Y}$ and ϕ denotes the normal probability density function. The expected jump size (9) is calculated by numerical integration of $w \mapsto w \cdot f_W(w|\Delta\hat{\kappa}_{t_Y})$, where the density f_W is obtained analogously to (11).

The described procedure aims at reverting any mortality shock at time t_Y before using it as the forecast jump-off year, which should diminish the persistence of shocks in forecasts of the jump model.

Regime switching. We consider the regime switching (RS) approach proposed by Milidonis *et al.* (2011). Its basic assumption is that the distribution of $\Delta\hat{\kappa}_t$ is governed by an unobservable binary Markov process ρ_t with values in $\{1, 2\}$. We expect that there will be one regime corresponding to normal mortality development and another regime corresponding to short periods of shock events. This would be in line with the findings of Lemoine (2015) on French data, who detect a high-volatility regime after the Second World War and a low-volatility regime on more recent data.

We initialize parameters before model calibration in such a way that state 1 should correspond to the normal regime and state 2 should correspond to the shock regime with higher expected drift of $\Delta\hat{\kappa}_t$ or significantly higher volatility.³ More precisely, we assume

$$\Delta\hat{\kappa}_t = \begin{cases} e_t^1 \sim \mathcal{N}(\mu, \sigma^2), \text{ i.i.d. if } \rho_t = 1, \\ e_t^2 \sim \mathcal{N}(m, s^2), \text{ i.i.d. if } \rho_t = 2, \end{cases} \tag{12}$$

where e_t^1 and e_t^2 are independent. Transition between the two regimes is described by the right stochastic matrix

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}, \tag{13}$$

where $p_{jk} := \mathbb{P}(\rho_{t+1} = k | \rho_t = j)$. This implies $p_{12} = 1 - p_{11}$ and $p_{22} = 1 - p_{21}$.

The model is calibrated by maximum likelihood estimation as detailed in Milidonis *et al.* (2011). Forecasts are obtained via Monte Carlo simulation (with $J = 10^6$ paths) from (12). Similarly to the jump model, we need an estimate of the hidden state process ρ_{t_Y} to obtain forecasts for ρ_{t_Y+1} and then $\Delta\hat{\kappa}_{t_Y+1}$. We set $\rho_{t_Y}^a := 1$ on Monte Carlo paths $a = 1, \dots, \pi J$ and $\rho_{t_Y}^a := 2$ on Monte Carlo paths $a = \pi J + 1, \dots, 10^6$, where

$$\pi := \mathbb{P}(\rho_{t_Y} = 1 | \Delta\hat{\kappa}_{t_Y}, \dots, \Delta\hat{\kappa}_{t_2}), \tag{14}$$

³Of course, this cannot be guaranteed, for example, when no larger shocks have occurred in a population, but we have found that it usually works well in practice.

see Milidonis *et al.* (2011, Appendix A) for details on how to estimate this probability. This approach ensures that appropriate regime transition probabilities are used when forecasting $\hat{\kappa}_{t_Y+1}$. Any mortality shock at time t_Y would, however, persist in forecasts because the jump-off value $\hat{\kappa}_{t_Y}$ is used without modification.

Summarizing, in the lognormal, POT, and RS models, the unmodified period effect $\hat{\kappa}_{t_Y}$ is used as a jump-off value for forecasting so that a shock in t_Y will persist in forecasts, at least temporarily. This could be amended by using a more robust jump-off value such as the average or even the median of $\hat{\kappa}_{t_Y}, \hat{\kappa}_{t_Y-1}, \dots, \hat{\kappa}_{t_Y-q}$ for some $q \in \mathbb{N}$ [Booth *et al.* (2006)]. Alternatively, the shocked period effect at t_Y could be “corrected” by estimating the impact of the mortality shock and subtracting it from $\hat{\kappa}_{t_Y}$ before using this as an input for forecasting $\hat{\kappa}_{t_Y+1}$.⁴ Both these adjustments would only change the intercept of the mortality trend, not its slope and also not the size of the estimated prediction uncertainty. We leave the identification and implementation of a suitable approach for the lognormal, POT, and RS models for future work.

2.4. A modified shock model calibration procedure

Chen and Cummins (2010), Chen and Cox (2009), and Milidonis *et al.* (2011) all propose to calibrate their models on a long observation period of mortality data. On the one hand, this makes sense since the models have to see sufficiently many data to get a good understanding of how extreme observations look like. On the other hand, the parameters $\alpha_x, \beta_x, \mu,$ and σ are assumed to be constant over time. Such an assumption of time invariance over long calibration windows has been proven wrong for many populations in the literature, see in particular Carter and Prskawetz (2001); Lee and Miller (2001); Li *et al.* (2013) for β_x and Booth *et al.* (2002); Sweeting (2011); Li *et al.* (2011) for μ .

Therefore, we propose to use a reduced calibration data set to ensure that the assumptions of the LC model are met. More precisely, we calibrate the LC and RWD parameters $\alpha_x, \beta_x, \mu,$ and σ on a shorter time period than the “shock parameters” θ , where $\theta = (z, \xi, \zeta)$ for the POT model, $\theta = (m, s, p)$ for the jump model, and $\theta = (m, s, p_{11}, p_{21})$ for the RS model. We modify the calibration for the POT, jump, and RS approaches introduced in section 2.3 as follows:

1. Calibrate an LC model on all available training data, e.g., on all the years from 1900 to 2020 to obtain parameters $\alpha_x^{\text{long}}, \beta_x^{\text{long}},$ and κ_t^{long} .
2. Calibrate a POT/jump/RS model on the estimated period effects $\hat{\kappa}_t^{\text{long}}$ from step 1 to obtain θ^{long} (shock parameters).
3. Calibrate another LC model on a reduced training data set with $t \in \mathcal{T}^{\text{red}}$, e.g., only on the years from 1991 to 2020 to obtain parameters $\alpha_x, \beta_x,$ and κ_t .
4. Choose μ and σ such that they maximize the profile likelihood (defined analogously as in (7)) for the observations $\Delta \hat{\kappa}_t, t \in \mathcal{T}^{\text{red}}$, with fixed $\theta = \theta^{\text{long}}$.⁵

The procedure yields a shock model with LC parameters $\alpha_x, \beta_x, \kappa_t,$ and period effect time series parameters $\mu, \sigma, \theta^{\text{long}}$. This can be interpreted as a model on \mathcal{T}^{red} which

⁴We do this for the jump model because such a correction is part of its model structure (8).

⁵For the POT model, the threshold u is calculated on the reduced data set \mathcal{T}^{red} as well by setting it to the $z\%$ quantile of the observations $(\Delta \hat{\kappa}_t)_{t \in \mathcal{T}^{\text{red}}}$. This ensures that $z\%$ of the observations to which μ and σ are calibrated are below the threshold u .

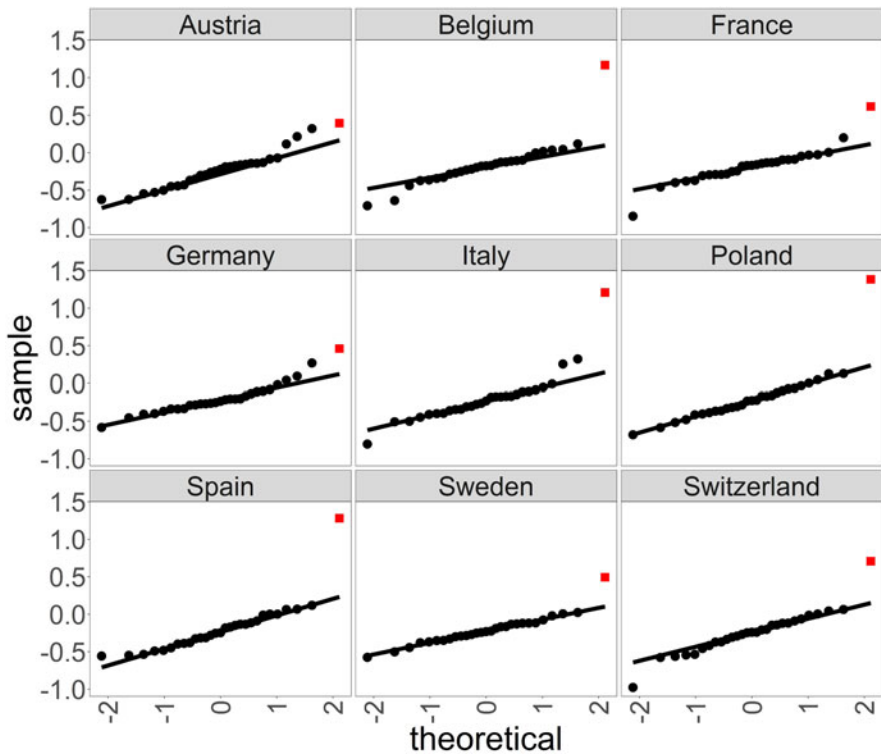


Figure 1. Quantile–quantile plots comparing the empirical distribution of $\Delta\hat{\kappa}_{1992}, \dots, \Delta\hat{\kappa}_{2019}$ (black circles) and $\Delta\hat{\kappa}_{2020}$ (red square) to a normal distribution.

incorporates the shock parameters θ^{long} as prior knowledge. They are calibrated on the whole available data set in the preliminary step 2, whereas all remaining parameters are calibrated on a shorter data set (steps 3 and 4) to ensure that the assumptions underlying the LC model are not violated. In our empirical studies, we use the last 30 years of data to calibrate the LC parameters as well as the drift and volatility of the period effects.

3. Empirical results

3.1. Checking the normal distribution assumption

Some of the models introduced in section 2.3 deviate from the normal distribution assumption for the period effect increments $\Delta\hat{\kappa}_t$. To evaluate whether this is justified as a consequence of the 2020 mortality shock, we compare their empirical distribution to the normal distribution with quantile–quantile plots in Figure 1. It is evident that the normal distribution assumption is broadly justified for the years up to 2019, but the point corresponding to 2020 is clearly visible as an outlier for several populations. Certainly, based on this observation, one could argue for using an outlier adjustment method to exclude these data. However, this is questionable since one cannot rule out the possibility of future pandemics or other shock events occurring.

We apply the Shapiro–Wilk test [Shapiro and Wilk (1965)], a statistical test to check whether the observed period effect increments have been generated from a normal distribution.⁶ Table 2 compares its p -values when calibrating the LC model on years up to 2019 and on years up to 2020. Unsurprisingly, they generally decrease, indicating a tendency to reject the normality assumption when 2020 is included in the calibration period.

We have performed multiple (54) statistical tests here, which means some adjustment procedure has to be applied to the p -values shown in Table 2 in order to keep the family-wise error rate below a desired significance level. Using the Bonferroni–Holm correction [Holm (1979)] we find that normality of the period effect increments is never rejected at a 5% significance level for models calibrated on 1990–2019 data, whereas it is rejected for 12 out of 27 populations for models calibrated on 1991–2020 data. These are the male, female, and total populations of Belgium, Poland, and Spain as well as French males, Italian males, and the total Italian population.

3.2. Evaluating our shock model calibration procedure

We evaluate the adjusted calibration procedure we have proposed in section 2.4 via a backtest. For this, we calibrate

1. all the parameters on all available data up to 2010 (“long”),
2. only the shock parameters on all available data up to 2010, the remaining LC parameters on years 1981–2010 as described in section 2.4 (“short”).

Then, we use both models to produce 1- to 10-year point forecasts $\hat{m}_{x,t}^i$ for the years $t = 2011, \dots, 2020$. These are compared to the ground truth $m_{x,t}^i$ by calculating the median average percentage error

$$\text{MdAPE}(t) := \text{median}_{x \in \mathcal{X}, i \in \mathcal{P}} \frac{|\hat{m}_{x,t}^i - m_{x,t}^i|}{m_{x,t}^i} \cdot 100\%. \quad (15)$$

As we observe from Figure 2, MdAPEs are clearly lower when a shorter time period is used for calibrating the LC model parameters and the normal mortality drift μ and volatility σ . The shock year 2020 is an exception, indicating that calibrating these parameters over an extensively long time horizon has made them overly pessimistic. Indeed, we find that using a smaller calibration period often significantly reduces μ or σ .

As reliable point forecasts are the most important requirement for a mortality model, we conclude that the proposed adjustment to the calibration is highly recommendable. We will use it throughout this section.

3.3. Shock probability and severity

The jump and the RS model provide information on the probability and the size of mortality shocks in the observed data. Figure 3 shows the period effect increments and probability of being in the normal state of the RS model. This probability is

⁶Other statistical tests could be applied as well. For instance, Lemoine (2015) proves that $\Delta \hat{\kappa}_t$ exhibits non-Gaussian properties on French data by applying the Jarque–Bera test.

Table 2. *p*-values of Shapiro–Wilk normality tests for all populations and jump-off years 2019 (models calibrated on 1990–2019) and 2020 (models calibrated on 1991–2020)

Population	2019	2020
Austria_F	0.624	0.768
Austria_M	0.186	0.025
Austria_T	0.497	0.128
Belgium_F	0.280	0.000
Belgium_M	0.084	0.000
Belgium_T	0.191	0.000
France_F	0.004	0.030
France_M	0.044	0.001
France_T	0.013	0.004
Germany_F	0.414	0.346
Germany_M	0.622	0.004
Germany_T	0.552	0.020
Italy_F	0.187	0.007
Italy_M	0.322	0.000
Population	2019	2020
Italy_T	0.218	0.000
Poland_F	0.402	0.000
Poland_M	0.686	0.000
Poland_T	0.708	0.000
Spain_F	0.828	0.000
Spain_M	0.612	0.000
Spain_T	0.286	0.000
Sweden_F	0.685	0.585
Sweden_M	0.201	0.008
Sweden_T	0.818	0.011
Switzerland_F	0.891	0.389
Switzerland_M	0.295	0.042
Switzerland_T	0.113	0.029

mostly close to one in all nine countries from 1975 to 2020. In 2020, a change of regime is experienced by six of them.

Table 3 displays the probabilities and expected sizes of a mortality jump conditional on the observed period effect change in 2020 according to the jump model (8). Apart from one exception, jump probabilities are larger than 60%. However, jump severities differ significantly by country. Furthermore, we have observed that in order to obtain

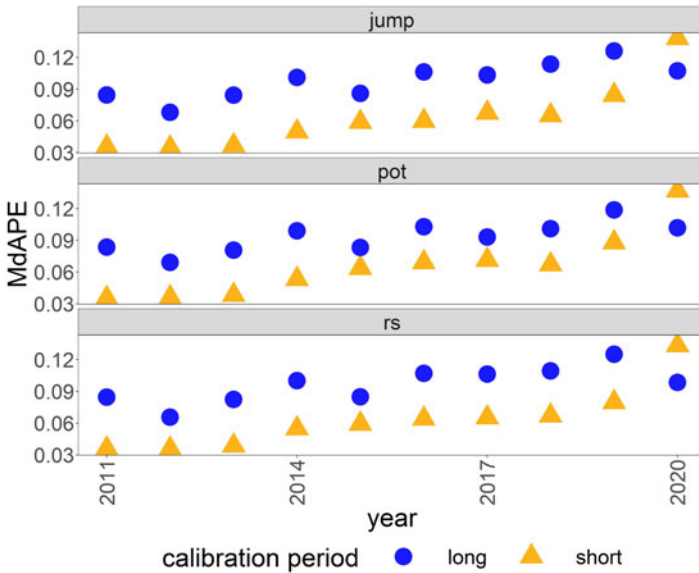


Figure 2. Yearly MdAPE between 2011 and 2020 for shock models with $\alpha_s, \beta_s, \mu, \sigma$ calibrated on data between 1981 and 2010 (“short”) or all available data (“long”).

sensible estimates, a sufficiently long history of mortality data is necessary. For some populations with relatively short mortality history (e.g., German males or Polish males), estimated jump probabilities are close to 100%, but the expected jump size is 0 because not enough historical mortality jump observations are available to calibrate an informative jump severity distribution.

3.4. Model parameters

We expect the additional parameters introduced by the POT, jump, and RS models to have a stabilizing effect on μ and σ , the drift and volatility of the period effect increments when mortality rates are normal. The lacking robustness of the RWD drift parameter is a well-known issue even under normal mortality and has been termed recalibration risk by Cairns (2013).

Figure 4 shows μ and σ of the shock models, comparing them to the RWD over two different calibration periods. When including the year 2020 in the calibration data set, μ and σ under the RWD clearly increase for all the countries. Considering the other models, these parameters are indeed less influenced by the extreme behavior of 2020 and there are only a few larger changes. For example, the volatility for Switzerland under the RS model increases. This is not surprising since the mortality shock in this country has not been large enough to lead to a regime change (see Figure 3), so that the observation of 2020 is used to calibrate μ and σ , increasing both.

3.5. Comparing the quality of fit

As all our models are calibrated via maximum likelihood estimation, a natural way of comparing their in-sample behavior consists in comparing their likelihoods. To

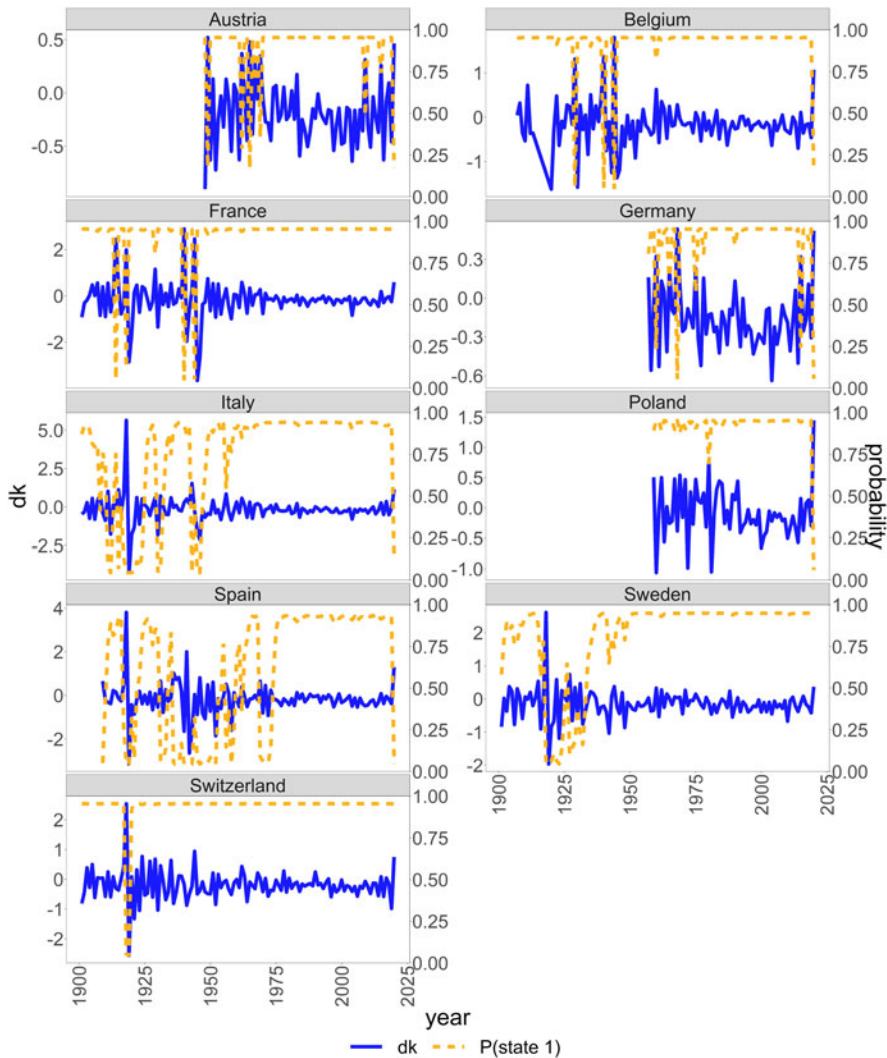


Figure 3. Period effect increments $\Delta\hat{\kappa}_t$ (blue, solid; left axis) and probability $\mathbb{P}(\rho_t = 1|\Delta\hat{\kappa}_t, \dots, \Delta\hat{\kappa}_2)$ of the “normal” state in the RS model (orange, dash; right axis).

account for the different number of parameters we consider the Bayesian information criterion

$$\text{BIC} := -2L_{\max} + \log(n_{\text{obs}}) \cdot n_{\text{par}}, \tag{16}$$

where L_{\max} is the model log-likelihood, n_{obs} the number of observations, and n_{par} the number of free parameters [see Cairns *et al.* (2009) for a previous application of the BIC to compare mortality models]. This penalizes the likelihood, giving preference to more parsimonious models. The likelihood and the number of observations are

Table 3. Conditional jump probability $\mathbb{P}(N_{2020} = 1|\Delta\hat{\kappa}_{2020})$ and expected jump size $\mathbb{E}(W_{2020}|\Delta\hat{\kappa}_{2020})$, see (10) and (9)

Country	Jump probability[%]	Expected jump size
Austria	62.7	0.120
Belgium	60.9	0.540
France	63.1	0.510
Germany	100.0	0.538
Italy	61.9	0.853
Poland	100.0	1.246
Spain	67.1	0.889
Sweden	27.9	0.168
Switzerland	77.3	0.725

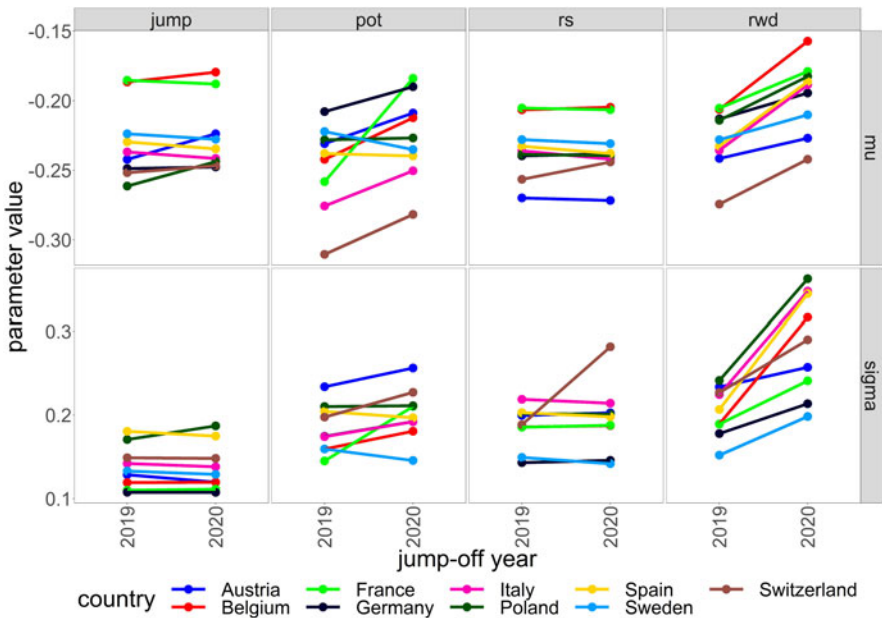


Figure 4. Model parameters μ and σ calibrated on the years 1990–2019 or 1991–2020 of the jump, peaks-over-threshold (pot), regime switching (rs), and random walk with drift (rwd) models.

calculated on the 30 years of data on which the LC parameters and the period effect drift and volatility are calibrated. For the shock models, we are actually including more information in the model by the calibration procedure described in section 2.4. We account for this by including the shock parameters θ in the total number of free parameters. This means we fully penalize for the increased model complexity, although these parameters are fixed in a first calibration step and the remaining

parameters are calibrated conditional on θ . Admittedly, this is a somewhat non-standard application of the BIC, but the results can nevertheless be assumed to give a good indication as to the balance of quality of fit and parsimony achieved by the models.

The BIC values of our models per country are displayed in Figure 5. The RWD mostly does not fit well, regardless of whether an underlying normal or lognormal distribution is assumed. In some cases, using a more general ARIMA model leads to better performance, but apart from Austria and Switzerland this is not optimal, either. The intervention model approach achieves low BIC values but is conceptually questionable due to its model structure, which does not allow for the future occurrence of non-normal mortality shocks. The shock models are more heavily penalized due to their higher number of parameters. The RS and jump models only in a few cases achieve a decrease in BIC compared to the RWD. The POT model has low BIC for all countries, which shows that the POT-based model adjustment improves the quality of fit even after accounting for the increased model complexity.

3.6. Backtest

In section 3.2, we have already performed a backtest to determine the *calibration method* for the POT, jump, and RS models, and we have found that the approach proposed in section 2.4 leads to better point forecasts. Here, using this new calibration method, we do another backtest to evaluate how the different *models* listed in Table 1 would have performed in the past.⁷ Analogously as in section 3.2, we calibrate the models on data from 1981 to 2010 (except for the shock parameters, which are also calibrated on all available data before 1981) and evaluate them on 2011–2020. For further details on backtesting methods for stochastic mortality models, we refer to Dowd *et al.* (2010), where the approach we take here is termed an expanding horizons backtest.

In Figure 6, we display the point forecast errors measured by MdAPE as defined in (15). We observe an increase over time, which is plausible as mortality rates further in the future are harder to predict and forecasting errors tend to accumulate over time, with a peak in the shock year 2020. Differences between the models are negligible.

The situation is different for interval forecasts as measured by the prediction interval coverage probability

$$PICP(t) := \frac{1}{|\mathcal{X}| \cdot |\mathcal{P}|} \sum_{x \in \mathcal{X}} \sum_{i \in \mathcal{P}} 1_{\{m_{x,t}^i \in [\hat{m}_{x,t}^{i,lower}, \hat{m}_{x,t}^{i,upper}]\}}, \tag{17}$$

where 1 denotes an indicator function and $[\hat{m}_{x,t}^{i,lower}, \hat{m}_{x,t}^{i,upper}]$ is the 95% prediction interval implied by the model, which is calculated by Monte Carlo simulation or explicit formulae, if available [see section 2 and Schnürch *et al.* (2022)]. It seems that the RWD as well as the more general ARIMA model and the intervention model underestimate the uncertainty in their forecasts, in particular for the shock year 2020. Assuming a lognormal instead of a normal distribution does not lead to significant changes of the PICP. The jump model has the highest PICP initially, but

⁷We exclude the best estimate approach from the backtest because it only makes sense under the condition that we know if and when there is a shock in the data, which is generally not the case for the data we calibrate our models on for this backtest.

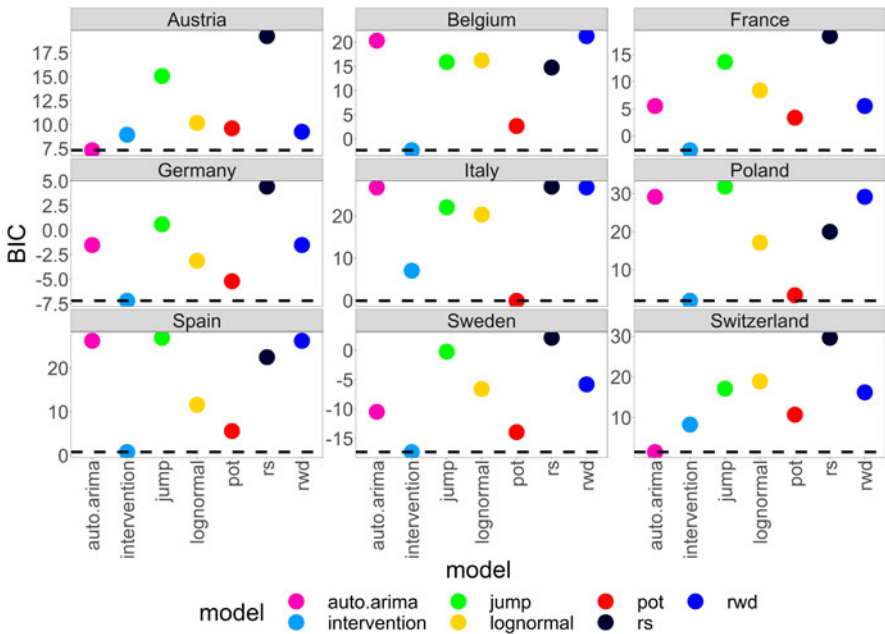


Figure 5. BIC for models calibrated on data from 1991 to 2020. The black dashed line is at the level of the best-fitting model.

it then decreases over time. This is due to its model structure and will be discussed further in section 3.7. The POT model and particularly the RS approach outperform the other methods for longer forecasting horizons. However, for the RS model this comes at the cost of very wide prediction intervals in some cases so that its intervals are more reliable but less informative.

The PICP is well below 95% for all models and time points. This is due to the fact that we ignore some sources of uncertainty in our consideration. For example, the prediction intervals could be extended to include parameter uncertainty via a bootstrapping approach [Koissi *et al.* (2006)].

As a robustness check, we have repeated the above evaluation for 99.5% prediction intervals and have found that the results are qualitatively similar.

3.7. Period effect forecasts

In this section, we illustrate the differences in period effect forecasts between the standard choice for a time series model, which is the RWD, and selected alternatives. In Figure 7, we display the observed period effects $\hat{\kappa}_t$, $t = 1991, \dots, 2020$, and their point and 95% interval forecasts according to the RWD, jump, POT, and RS models. For Sweden, a population with rather small estimated jump probability and size (see Table 3), all point forecasts use a jump-off value equal or close to $\hat{\kappa}_{2020}$, and mostly agree in their trend, staying relatively similar over time. Interval forecasts differ significantly: the jump model has the most narrow intervals, followed by the standard RWD approach. The POT model has almost the same lower interval bounds as the RWD, but its upper bounds are more conservative. The RS model has

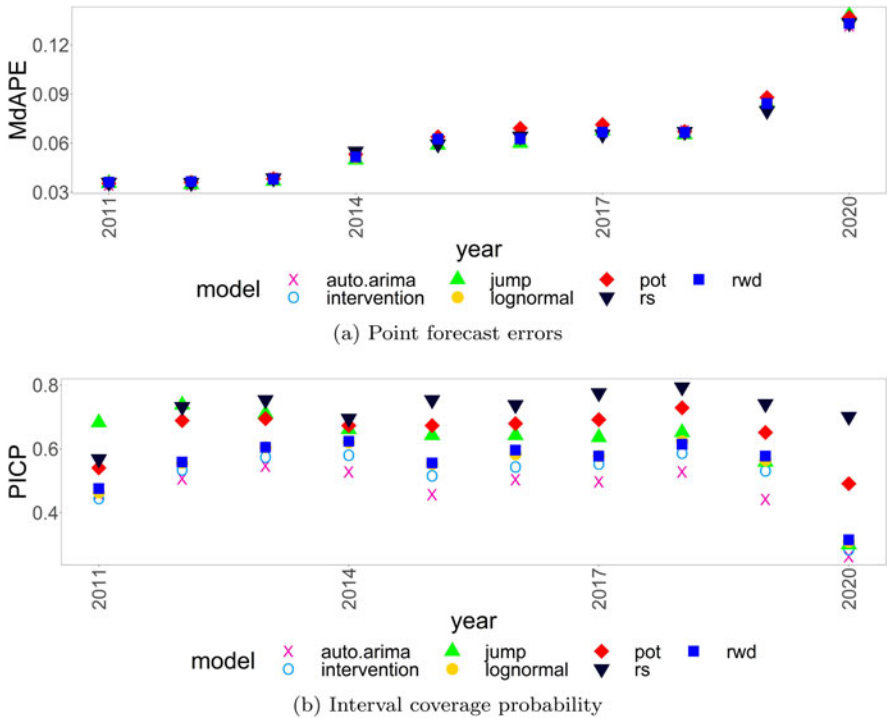
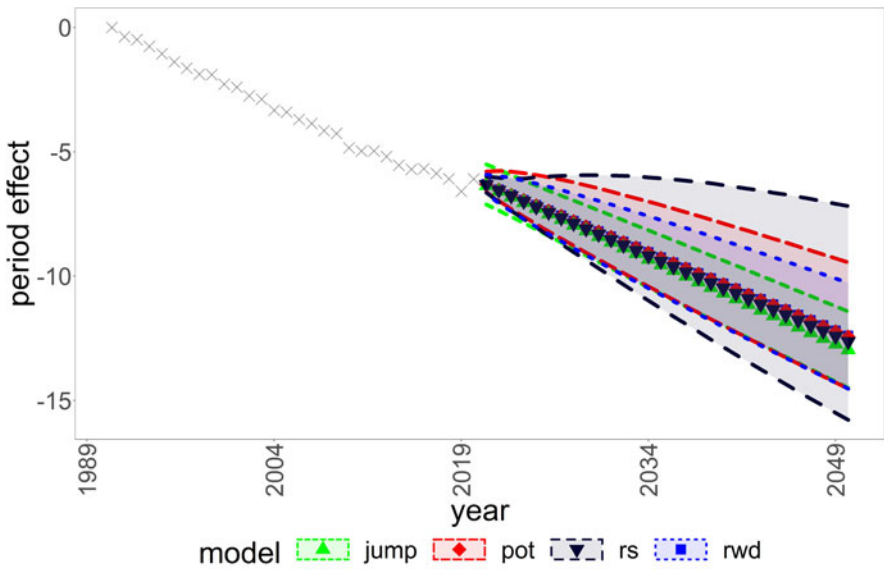


Figure 6. Point and interval forecast performance measures from 2011 to 2020 (models calibrated on data from 1981 to 2010). (a) Point forecast errors, (b) interval coverage probability.

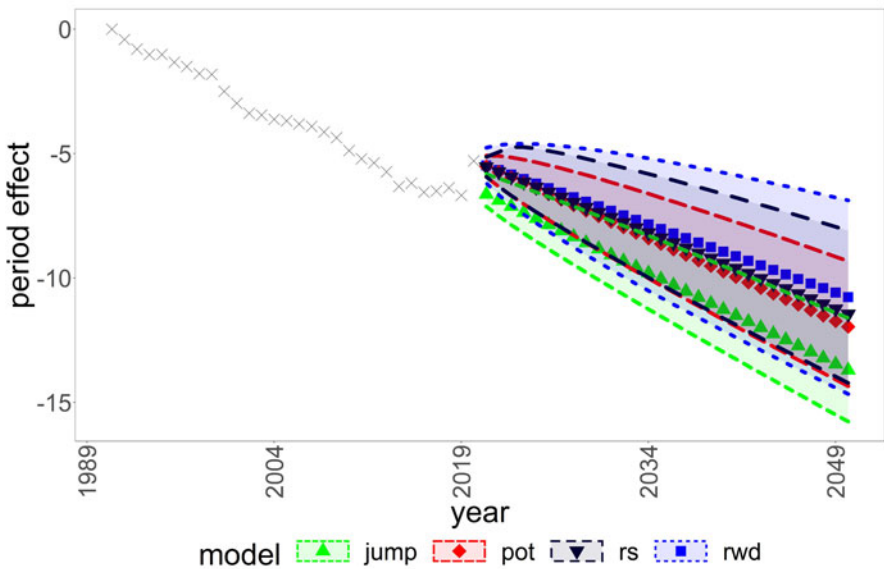
the widest intervals except in the first few years. This is probably due to its two quite distinct regimes ($\mu = -0.23, \sigma = 0.14$ vs. $m = -0.10, s = 0.96$).

In contrast, point forecasts visibly differ between models for Poland, which has experienced a relatively large shock to its 2020 period effect. In particular, the jump model reverts much of this shock before forecasting so that its central projection is more optimistic than that of the other models. Still, POT and RS predict a stronger decreasing trend than RWD. Prediction interval width is largest for the RWD, which is probably due to the fact that the 2020 mortality shock leads to a large increase in period effect volatility. The other models are able to (partially) capture this volatility in their shock parameters (s in the jump and RS model, the GPD parameters in the POT model), which usually have a smaller impact on prediction intervals than volatility at “normal” times.

For an easier visual comparison of all models and countries, we display point and interval forecasts in the last forecast year, $\hat{\kappa}_{2050}$, in Figure 8. Allowing for a more general ARIMA process (auto.arima) than the RWD tends to decrease point forecasts and leads to more narrow prediction intervals for some countries (e.g., Belgium), which is possibly due to ARIMA models generally not putting as much emphasis on the last observed value $\hat{\kappa}_{2020}$ as the RWD. Modeling the period effect with a lognormal distribution leads to almost identical point forecasts as the RWD, whereas the prediction intervals have a tendency to be slightly narrower. Unsurprisingly, the



(a) Sweden



(b) Poland

Figure 7. Observed period effects $\hat{\kappa}_t$, $t = 1991, \dots, 2020$, as well as point and 95% interval forecasts $\hat{\kappa}_t$ of selected models, $t = 2021, \dots, 2049$. Intervals obtained by closed-form expressions (RWD) or Monte Carlo simulation (jump, POT, RS). (a) Sweden, (b) Poland

intervention and best estimate approaches are more optimistic than the RWD, with lower point forecasts in all countries. Their prediction intervals are narrower because they ignore the extreme observation of 2020.

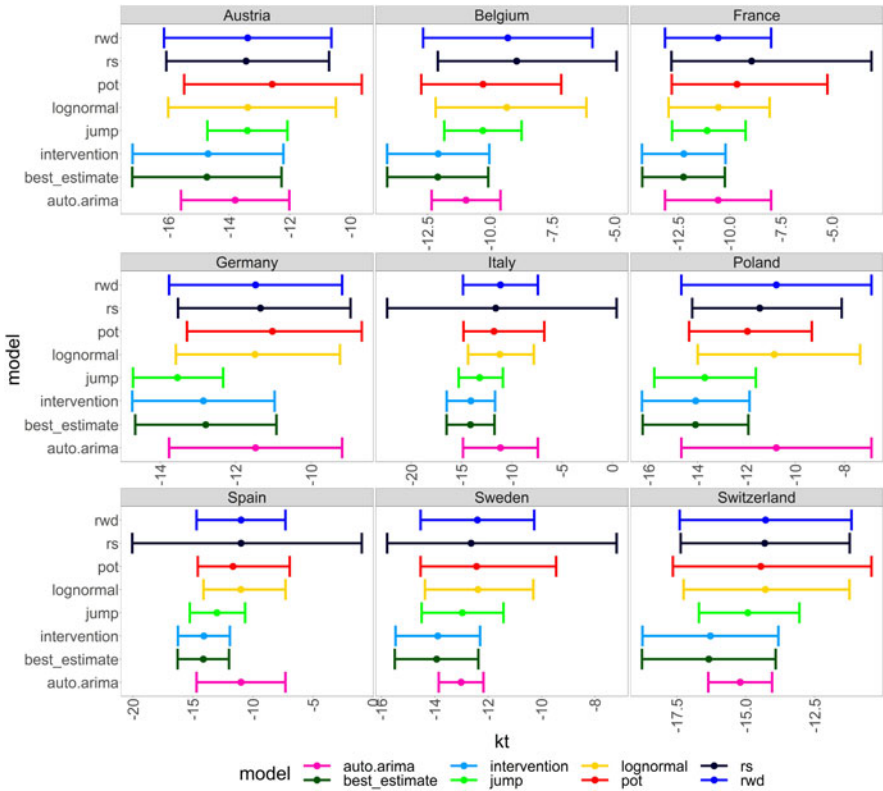


Figure 8. Country-specific period effect point and 95% interval forecasts for the year 2050, $\hat{\kappa}_{2050}$, comparing different time series forecasting methods. Intervals obtained by closed-form expressions or Monte Carlo simulation.

The jump model usually has lower point forecasts than the RWD because jumps vanish after one period, which is implemented by reducing $\hat{\kappa}_{2020}$ by the expected jump size before using it as the forecast jump-off value, see section 2.3. The model also tends to have narrow prediction intervals. A look at the parameters indicates that it separates the period effect increment distribution into two components: a normal component with low drift μ and low volatility σ and a shock component with higher, usually positive drift m and volatility $s > 0$. The influence of the shock component for long-term forecasts is small compared to the influence of the normal component because the jumps are modeled as transitory, so they do not have a lasting impact. Therefore, the reduction in normal volatility σ compared to an ordinary RWD model can lead to lower prediction uncertainty.

The RWD, POT, and RS point forecasts are broadly similar and there is no country-independent tendency regarding their order. For a few countries (e.g., Italy), the RS prediction intervals are substantially wider than the RWD intervals, which is due to the high estimated volatility (e.g., $s = 1.7$) in the shock regime combined with the non-negligible probability of transitioning into this regime from the normal state (e.g., $p_{12} = 4.9\%$). It is visible that POT and RS put higher emphasis on (upward) mortality shocks because the prediction intervals are often asymmetric with a wider upper part.

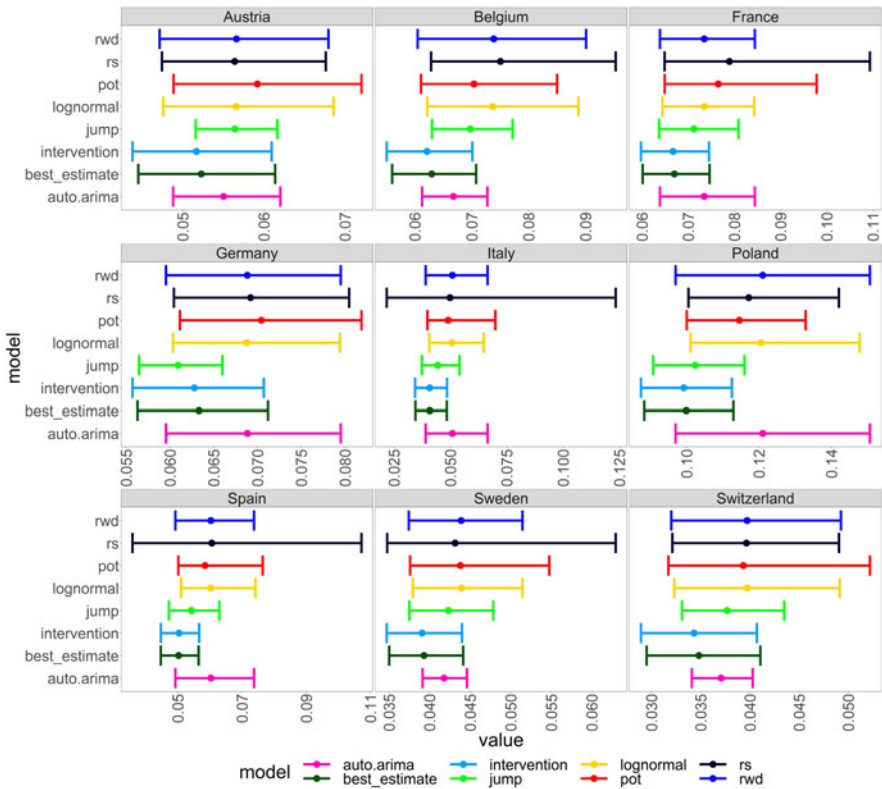


Figure 9. Country-specific life insurance values $A_{35:30}$ (2021) based on point and interval death rate forecasts, comparing different time series forecasting methods. Discount factor $\nu = \frac{1}{1.005}$. The bars are obtained by inserting the 95% prediction interval bounds for the death rates into (18).

3.8. Life insurance and annuity values

The mortality forecasts of our models can be summarized and we can obtain a better understanding of model risk by calculating life insurance and annuity values [see Richards and Currie (2009); Schnürch *et al.* (2022)]. For this, we assume a constant yearly discount factor $\nu = \frac{1}{1.005}$, which is a simplification allowing us to focus on the differences in mortality forecasts. As a typical example we consider a term assurance issued to a life aged $x = 35$ at the start of year $t = 2021$ which runs for $n = 30$ years and pays an amount of 1 at the end of the year if the life has died within this year. Its present value is given by

$$\begin{aligned}
 A_{x:\overline{n}|}(t) &:= \sum_{s=0}^{n-1} \nu^{s+1} {}_s p_{x,t} (1 - p_{x+s,t+s}) \\
 &\approx \sum_{s=0}^{n-1} \nu^{s+1} \exp\left(-\sum_{j=0}^{s-1} m_{x+j,t+j}\right) (1 - \exp(-m_{x+s,t+s})).
 \end{aligned}
 \tag{18}$$

Here, we have used the approximation $p_{x,t} \approx \exp(-m_{x,t})$ [Pitacco *et al.* (2008), Chapter 2.3].

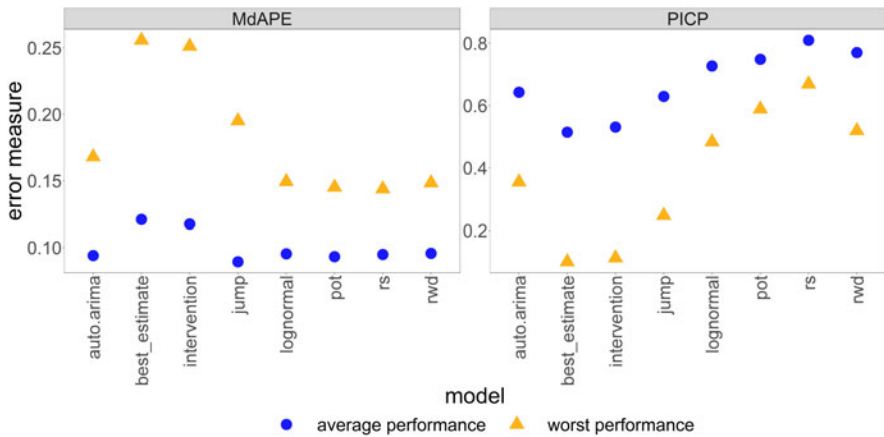


Figure 10. Point and interval forecast errors over different future mortality scenarios, average and worst case. Model calibrated on data from 1991 to 2020.

Similarly, we determine the present value of a temporary annuity-immediate for a life aged $x = 65$ at the start of year $t = 2021$ running for $n = 30$ years, denoted by $a_{x:\overline{n}|}(t)$. We calculate intervals for annuity and life insurance values by inserting the 95% prediction interval bounds for the death rates into the annuity and life insurance valuation formulae.⁸

Figure 9 shows life insurance values implied by the different models. Annuity values are shown in Figure 12 (Appendix B). Observations from these figures are mostly analogous to the ones about period effect forecasts in section 3.7.

3.9. Model performance under different future scenarios

It would be useful to get some impression of the future forecasting accuracy of the different models. This is complicated by the fact that the occurrence of mortality shocks and the further course of the COVID-19 pandemic are very hard to predict. While most scientists believe that a transition to an endemic phase is the most likely trajectory and a worldwide elimination of the virus is considered almost impossible [Scudellari (2020); Phillips (2021)], it is unclear how long such a transition will take and how severe the disease will continue to be.

Telenti *et al.* (2021) describe three future scenarios: (i) ongoing high levels of infection and severe disease due to the inability to gain rapid control over COVID-19, (ii) transition to an endemic seasonal disease similar to influenza, which has a yearly global death toll approximately between 250000 and 500000, (iii) transition to an endemic disease similar to other existing human coronavirus infections with significantly lower impact on population health and mortality. This

⁸Our approach illustrates the possible range of annuity and life insurance values assuming that mortality rates are consistently at the boundaries of their 95% prediction intervals. Point and interval forecasts of present values could be obtained by Monte Carlo simulation, which however requires a lot of computational effort. Additionally, we have verified numerically that our approach yields only slightly different values compared to Monte Carlo simulation for the data under consideration.

last state could, however, take decades to reach or may not occur at all. They further note that the probability of occurrence is hard to quantify for any of these scenarios due to material gaps in current scientific knowledge.

Another aspect influencing the assessment of future mortality development is the possibility that SARS-CoV-2 might reoccur years after the alleged end of the COVID-19 pandemic, for instance by spilling back from animal reservoirs into the human population [Phillips (2021)]. Brüssow (2021) points out that COVID-19 displays clinical similarities to the 1889 Russian flu pandemic, which lasted for several years and then possibly reappeared around ten years later. As a related alternative, a new mortality shock might arise, for example, in the form of another pandemic caused by a different pathogen such as MERS-CoV: “the risk for MERS-CoV evolving into a more transmissible virus should not be underestimated” [Telenti *et al.* (2021)]. It is impossible to predict the exact time, but it is highly likely that further pandemics will emerge: “the question is not if, but when” [World Health Organization (2021)].

With these considerations in mind, we define five exemplary 30-year scenarios of future mortality with a focus on the development of COVID-19 and the occurrence of new mortality shocks. Detailed definitions of these scenarios are provided in [Appendix A](#). Generally, two of them are relatively optimistic (second wave in 2021 and normality afterwards; only short-term mortality impairments), while the others are more pessimistic (occurrence of another pandemic in 2034; occurrence of another pandemic in 2048; persistent consequences of the COVID-19 pandemic). Our scenarios for the future impact of COVID-19 on mortality are inspired by the ones considered by Continuous Mortality Investigation (2020). For the scenarios including the occurrence of a further pandemic, we have chosen the years 2034 and 2048 arbitrarily for illustrative purposes. Zhou and Li (2021) show how to include expert opinions about disease transmission and infection fatality rates on the short term and about the reoccurrence of a COVID-alike pandemic on the longer term in the definition of mortality scenarios.

In order to get a better impression of each model’s forecasting behavior, we evaluate its average and worst-case performance over our five scenarios. [Figure 10](#) shows the results of this evaluation.

The models usually achieve their worst performance in the most pessimistic scenario, which assumes a persisting interruption of mortality improvements. The best estimate and intervention models are the least robust with respect to this scenario and have the highest worst-case MdAPE and lowest worst-case PICP. The remaining models have similar point forecasts and, therefore, similar average and worst-case performances (the jump model is a slight exception). However, the prediction intervals of the POT and the RS model are more robust with respect to their worst-case errors compared to the RWD.

With a scenario-based approach, model choice can be grounded upon the beliefs of the modeler by assigning subjective probabilities to each of the scenarios, calculating a correspondingly weighted average of the error measures under each scenario and choosing the model which minimizes this weighted average. Of course, this requires sufficient information to estimate occurrence probabilities, which will only be obtained over time.

4. Conclusion

Depending on the considered population, the usual RWD is not always appropriate for modeling the LC period effects because it relies on a normal distribution assumption, which is not robust toward mortality shocks. In particular, the excess mortality induced by COVID-19 strongly influences point forecasts in the short term and interval forecasts even in the long term [Schnürch *et al.* (2022)].

In this work, we have demonstrated that simply switching to a different distribution such as the lognormal without structurally modeling mortality shocks does not sufficiently address this issue. General outlier adjustment methods such as an intervention model improve the fit and allow to keep the normal distribution assumption, but they have poorer backtest performance as regards prediction uncertainty quantification (section 3.6), and both their point and interval forecasts would be significantly inferior in an adverse mortality scenario (section 3.9).

We therefore recommend to not exclude the death rates of 2020 from the training data of the LC model nor to use an intervention model, unless mortality shocks are not deemed relevant for the application. Otherwise, the mixture model approach based on the POT method looks most appropriate in our empirical investigations. It yields a better fit as well as more stable and reliable interval forecasts than the RWD. Depending on the views of the modeler on future mortality trends, an additional adjustment to diminish the persistence of the 2020 shock in its forecasts may be advisable (see the remark at the end of section 2.3).

While being conceptually intriguing, the RS approach has rather high BIC values and yields very wide prediction intervals. The jump model leads to more narrow prediction intervals, which indicates that it in fact works as intended (see the explanation in section 3.7). It further contains an explicit adjustment of the forecast jump-off value to remove the impact of a shock, which is based on estimations of shock probability and size given the observed period effect change. However, its PICP decreases over time in the backtest and its quality of fit is suboptimal for most populations.

In any case, if using one of the three shock models, we have demonstrated in section 3.2 that it is clearly beneficial to calibrate the time series parameters relating to the shock events on a larger data set than the LC and the normal time series parameters (μ , σ) instead of calibrating all the parameters on the same large data set as proposed in the literature so far.

There are several directions for extensions and further research:

- In some countries, the COVID-19 pandemic might have affected old-age mortality more than young-age mortality. This is not reflected in the model approaches we have considered, as they only account for shocks which follow the same age pattern as general mortality improvements. One could address this by introducing shock-specific age effect parameters [Liu and Li (2015)]. However, this complicates calibration, among other reasons due to sparsity of data on mortality shock events. An alternative is to incorporate expert opinions, which could not only relate to age patterns of COVID-19 mortality but also to its future development and the occurrence of new pandemics [Zhou and Li (2021)]. This delicate issue is beyond the scope of this paper, and it certainly deserves further investigation.
- At the moment of writing, COVID-19 is expected to significantly affect mortality in 2021 as well, thus inducing a longer-lasting mortality shock. Such an effect is

explicitly modeled only by the RS model, where it corresponds to a longer-lasting stay in the shock regime. It could be incorporated in the jump model as well by modeling a serial correlation in the Bernoulli process governing the occurrence of jumps or even in the jump severities. Again, this would not be easy to calibrate. A simpler alternative is to reformat the data in two-year instead of one-year time intervals. This trivially allows to explicitly model shocks of up to two years. However, modeling in two-year intervals is somewhat unusual and reduces the available data by half.

- Instead of ordinary prediction intervals, we could consider highest-density regions as proposed by Hyndman (1995) to better account for multimodality of the jump and RS models. These regions can consist of several intervals, one for each mode, and therefore yield a more informative distributional prediction for future mortality rates.
- We have focused on the LC model, but the presented approaches are in principle applicable for other stochastic mortality models as well. It would be particularly interesting to consider models with more than one factor, which could require some multi-dimensional generalizations of the approaches from section 2.3.
- We have modeled all populations in isolation. One could instead consider a multi-population approach where all countries or groups of similar countries share the same shock parameters. Analogous ideas have been investigated in the literature for other parameters of stochastic mortality models [Li and Lee (2005); Kleinow (2015); Schnürch *et al.* (2021)], and a first proposal in this direction has been made by Zhou *et al.* (2013). However, in contrast to “normal” mortality, it is not implausible that the effects of mortality shocks can be strongly country-specific (e.g., German mortality in 2020 has been less affected by COVID-19 than in several neighboring countries). In fact, we have compared the parameters obtained over the considered populations, and they are usually quite different. Given that for some of the parameters, such as the RS probability p_{12} , even small differences can lead to large effects, we have refrained from implementing such an approach.

Acknowledgements. We would like to express our gratitude to the anonymous referee for the constructive comments. Furthermore, we thank Ralf Korn and Stephen Richards for helpful discussions. S. Schnürch is grateful for the financial support from the Fraunhofer Institute for Industrial Mathematics ITWM. T. Kleinow acknowledges financial support from the Actuarial Research Centre of the Institute and Faculty of Actuaries through the research program on “Modelling, Measurement and Management of Longevity and Morbidity Risk.”

References

- Balkema, A. A. and L. de Haan (1974) Residual life time at great age. *The Annals of Probability* 2(5), 792–804.
- Behrens, C. N., H. F. Lopes and D. Gamerman (2004) Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling: An International Journal* 4(3), 227–244.
- Booth, H., R. J. Hyndman, L. Tickle and P. de Jong (2006) Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research* 15289–310.
- Booth, H., J. Maindonald and L. Smith (2002) Applying Lee-Carter under conditions of variable mortality decline. *Population Studies* 56(3), 325–336.
- Box, G. E. P. and D. R. Cox (1964) An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 26(2), 211–243.

- Box, G. E. P. and G. C. Tiao (1975) Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association* 70(349), 70–79.
- Brüssow, H. (2021) What we can learn from the dynamics of the 1889 “Russian flu” pandemic for the future trajectory of COVID-19. *Microbial Biotechnology* 14(6), 2244–2253.
- Cairns, A. J. G. (2013) Robust hedging of longevity risk. *Journal of Risk and Insurance* 80(3), 621–648.
- Cairns, A. J. G., D. P. Blake, K. Dowd, G. D. Coughlan, D. P. Epstein, A. Ong and I. Balevich (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* 13(1), 1–35.
- Carter, L. R. and A. Prskawetz (2001) Examining structural shifts in mortality using the Lee-Carter method. MPIDR Working Paper WP 2001–2007, Max Planck Institute for Demographic Research, Rostock.
- Chen, H. and S. H. Cox (2009) Modeling mortality with jumps: applications to mortality securitization. *Journal of Risk and Insurance* 76(3), 727–751.
- Chen, H. and J. D. Cummins (2010) Longevity bond premiums: the extreme value approach and risk cubic pricing. *Insurance: Mathematics and Economics* 46(1), 150–161.
- Coles, S. G. and M. J. Dixon (1999) Likelihood-based inference for extreme value models. *Extremes* 2(1), 5–23.
- Continuous Mortality Investigation (2020) Considerations relating to COVID-19 for mortality and morbidity assumptions. Working Paper 139, Institute and Faculty of Actuaries.
- Dowd, K., A. J. G. Cairns, D. P. Blake, G. D. Coughlan, D. P. Epstein and M. Khalaf-Allah (2010) Backtesting stochastic mortality models. *North American Actuarial Journal* 14(3), 281–298.
- Engel, K. and S. Ziegler (2020) Pandora’s Box. A report on the human zoonotic disease risk in Southeast Asia with a focus on wildlife markets. Report, WWF.
- Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6(2), 65–70.
- Human Mortality Database (2021) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research, Rostock (Germany). Data downloaded on July 21 from www.mortality.org.
- Hyndman, R. J. (1995) Highest-density forecast regions for nonlinear and non-normal time series models. *Journal of Forecasting* 14(5), 431–441.
- Hyndman, R. J. and Y. Khandakar (2008) Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software* 27(3), 1–22.
- Kleinow, T. (2015) A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics* 63, 147–152.
- Koissi, M.-C., A. F. Shapiro and G. Högnäs (2006) Evaluating and extending the Lee–Carter model for mortality forecasting: bootstrap confidence interval. *Insurance: Mathematics and Economics* 38(1), 1–20.
- Lee, R. D. and L. R. Carter (1992) Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87(419), 659–671.
- Lee, R. D. and T. Miller (2001) Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography* 38(4), 537–549.
- Lemoine, K. (2015) Mortality regimes and longevity risk in a life annuity portfolio. *Scandinavian Actuarial Journal* 2015(8), 689–724.
- Li, J. S. -H. and W.-S. Chan (2005) Outlier analysis and mortality forecasting: the United Kingdom and Scandinavian countries. *Scandinavian Actuarial Journal* 2005(3), 187–211.
- Li, J. S.-H. and W.-S. Chan (2007) The Lee-Carter model for forecasting mortality, revisited. *North American Actuarial Journal* 11(1), 68–89.
- Li, J. S. -H., W. -S. Chan and S.-H. Cheung (2011) Structural changes in the Lee-Carter mortality indexes. *North American Actuarial Journal* 15(1), 13–31.
- Li, N. and R. D. Lee (2005) Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method. *Demography* 42(3), 575–594.
- Li, N., R. D. Lee and P. Gerland (2013) Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography* 50(6), 2037–2051.
- Liu, Y. and J. S.-H. Li (2015) The age pattern of transitory mortality jumps and its impact on the pricing of catastrophic mortality bonds. *Insurance: Mathematics and Economics* 64, 135–150.
- Mendes, B. V. d. M. and H. F. Lopes (2004) Data driven estimates for mixtures. *Computational Statistics & Data Analysis* 47(3), 583–598.

- Milidonis, A., Y. Lin and S. H. Cox (2011) Mortality regimes and pricing. *North American Actuarial Journal* 15(2), 266–289.
- Phillips, N. (2021) The coronavirus is here to stay—here’s what that means. *Nature* 590(7846), 382–384.
- Pickands, J. (1975) Statistical inference using extreme order statistics. *The Annals of Statistics* 3(1), 119–131.
- Pitacco, E., M. Denuit, S. Haberman and A. Olivieri (2008) *Modelling Longevity Dynamics for Pensions and Annuity Business*. Oxford: Oxford Univ. Press.
- R Core Team (2019) R: a language and environment for statistical computing. Vienna, Austria. www.R-project.org
- Regis, L. and P. Jevtić (2022) Stochastic mortality models and pandemic shocks. In Del Boado-Penas, M. C., Eisenberg, J., and Şahin Ş. (eds.), *Pandemics: Insurance and Social Protection*, Springer eBook Collection, pp. 61–74. Cham: Springer International Publishing.
- Ribatet, M. and C. Dutang (2019) POT: generalized Pareto distribution and peaks over threshold.
- Richards, S. J. and I. D. Currie (2009) Longevity risk and annuity pricing with the Lee-Carter model. *British Actuarial Journal* 15(2), 317–365.
- Scarrott, C. and A. MacDonald (2012) A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT* 10(1), 33–60.
- Schnürch, S., T. Kleinow and R. Korn (2021) Clustering-based extensions of the common age effect multi-population mortality model. *Risks* 9(3), 45.
- Schnürch, S., T. Kleinow, R. Korn and A. Wagner (2022) The impact of mortality shocks on modeling and insurance valuation as exemplified by COVID-19. *Annals of Actuarial Science* 16(3), 498–526.
- Scudellari, M. (2020) How the pandemic might play out in 2021 and beyond. *Nature* 584(7819), 22–25.
- Shapiro, S. S. and M. B. Wilk (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52(3–4), 591–611.
- Short Term Mortality Fluctuations (2021) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research, Rostock (Germany). Original input data; downloaded on July 21 from www.mortality.org.
- Sweeting, P. J. (2011) A trend-change extension of the Cairns-Blake-Dowd model. *Annals of Actuarial Science* 5(02), 143–162.
- Telenti, A., A. Arvin, L. Corey, D. Corti, M. S. Diamond, A. García-Sastre, R. F. Garry, E. C. Holmes, P. S. Pang and H. W. Virgin (2021) After the pandemic: perspectives on the future trajectory of COVID-19. *Nature* 596(7873), 495–504.
- Tuljapurkar, S., N. Li and C. Boe (2000) A universal pattern of mortality decline in the G7 countries. *Nature* 405(6788), 789–792.
- World Health Organization (2021) COVID-19 shows why united action is needed for more robust international health architecture.
- Zhou, R. and J. S.-H. Li (2021) A multi-parameter-level model for simulating future mortality scenarios with COVID-alike effects. Working Paper.
- Zhou, R., J. S.-H. Li and K. S. Tan (2013) Pricing standardized mortality securitizations: a two-population model with transitory jump effects. *Journal of Risk and Insurance* 80(3), 733–774.

Appendix A: Defining scenarios of future mortality

We define a 30-year mortality scenario by specifying improvement rates

$$I_{x,t} := \frac{m_{x,t-1} - m_{x,t}}{m_{x,t-1}} \quad (\text{A1})$$

for $x \in \mathcal{X}$ and $t = 2021, \dots, 2050$. We consider five scenarios:

- Second wave in 2021 and normality afterwards:

$$I_{x,2021} = 0, I_{x,2022} = 1 - \frac{1}{1 - I_{x,2020}}, I_{x,t} = I_{x,t-31} \text{ for } t = 2023, \dots, 2050. \quad (\text{A2})$$

In particular, we have $m_{x,2022} = m_{x,2019}$ and “normality” means that we observe exactly the same sequence of improvement rates between 2023 and 2050 as between 1982 and 2019.

- Short-term impairments:

$$I_{x,2021} = \frac{I_{x,1990} - I_{x,2020} - 0.04}{1 - I_{x,2020}}, I_{x,2021+h} = I_{x,1990+h} - 0.01(4 - h) \text{ for } h = 1, 2, 3, I_{x,t} = I_{x,t-31}$$

for $t = 2025, \dots, 2050$.

(A3)

In particular, we have $m_{x,2021} = (1.04 - I_{x,1990})m_{x,2019}$.

- Second wave in 2021 and normality afterwards until another pandemic inducing a larger shock than COVID-19 (with identical age structure) emerges in 2034:

$$I_{x,2021} = 0, I_{x,2022} = 1 - \frac{1}{1 - I_{x,2020}}, I_{x,t} = I_{x,t-31} \text{ for } t = 2023, \dots, 2033, I_{x,2034} = (2I_{x,2020})^-,$$

$$I_{x,2035} = 0, I_{x,2036} = 1 - \frac{1}{1 - I_{x,2034}}, I_{x,t} = I_{x,t-31} \text{ for } t = 2037, \dots, 2050.$$

(A4)

In particular, we have $m_{x,2022} = m_{x,2019}$ and $m_{x,2036} = m_{x,2033}$.

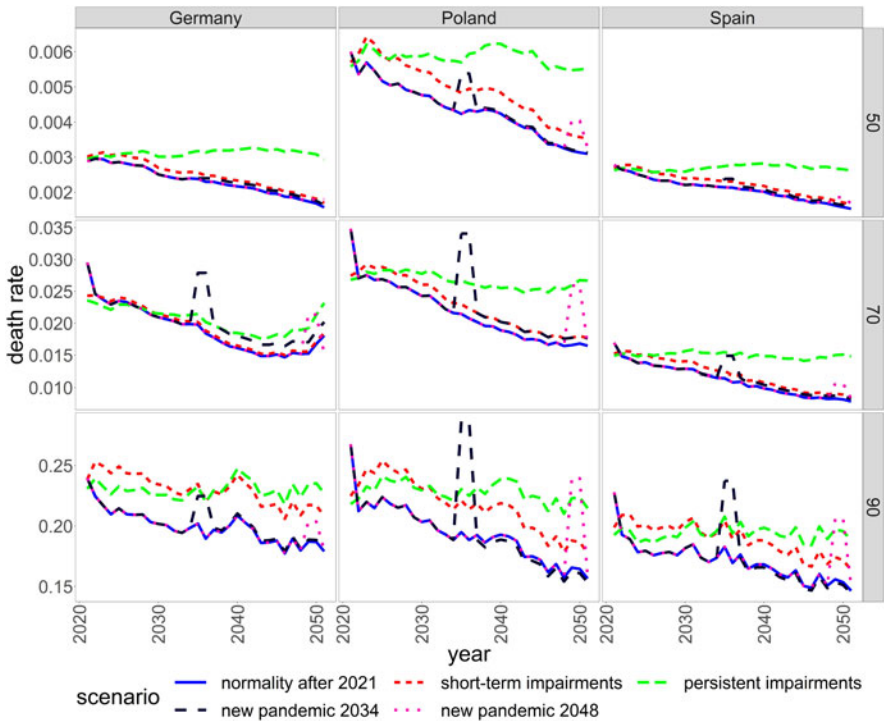


Figure A11. Development of mortality rates for the German, Polish, and Spanish populations in the age groups 50–54, 70–74, and 90+ between 2021 and 2050 under the five scenarios defined above.

- Second wave in 2021 and normality afterwards until another pandemic inducing a larger shock than COVID-19 (with identical age structure) emerges in 2048:

$$\begin{aligned}
 I_{x,2021} &= 0, I_{x,2022} = 1 - \frac{1}{1 - I_{x,2020}}, I_{x,t} = I_{x,t-31} \text{ for } t = 2023, \dots, 2047. I_{x,2048} = (2I_{x,2020})^-, \\
 I_{x,2049} &= 0, I_{x,2050} = 1 - \frac{1}{1 - I_{x,2048}}.
 \end{aligned}
 \tag{A5}$$

In particular, we have $m_{x,2022} = m_{x,2019}$ and $m_{x,2050} = m_{x,2047}$.

- Persistent reduction in improvement rates, leading to zero average improvement between 2022 and 2050,

$$I_{x,2021} = \frac{I_{x,1990} - I_{x,2020} - g}{1 - I_{x,2020}}, I_{x,t} = I_{x,t-31} - g \text{ for } t = 2022, \dots, 2050,
 \tag{A6}$$

where $g := \frac{1}{29} \sum_{t=1991}^{2019} I_{x,t}$. In particular, we have $m_{x,2021} = (1 + g - I_{x,1990})m_{x,2019}$.

Note that we do not make a statement about the probabilities of any of these scenarios becoming reality. The development of mortality rates under the five scenarios is shown in [Figure 11](#) for chosen countries and age groups.

Appendix B: Annuity values

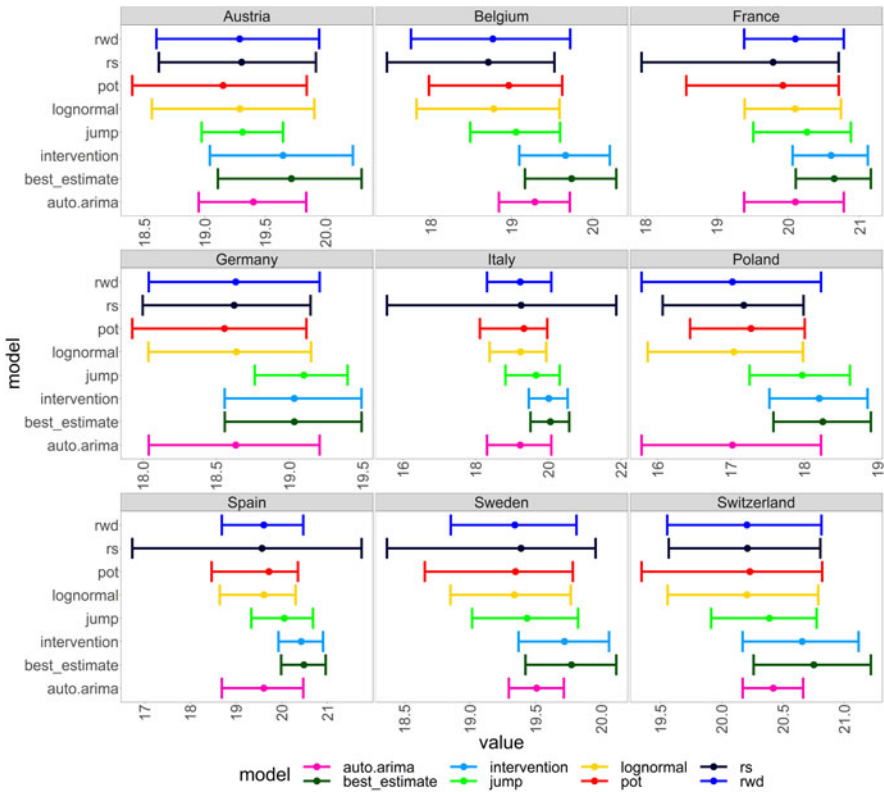


Figure B12. Country-specific annuity values $a_{65:\overline{30}|}(2021)$ based on point and interval death rate forecasts, comparing different time series forecasting methods. Discount factor $v = \frac{1}{1.005}$. The bars are obtained by calculating annuity values based on the 95% prediction interval bounds for the death rates.

Cite this article: Schnürch S, Kleinow T, Wagner A (2023). Accounting for COVID-19-type shocks in mortality modeling: a comparative study. *Journal of Demographic Economics* 89, 483–512. <https://doi.org/10.1017/dem.2023.9>