Projecting the future burden of cancer: Bayesian age–period–cohort analysis with integrated nested Laplace approximations

Andrea Riebler*[,1] and **Leonhard Held**[2]

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology, Alfred Getz vei 1, 7th floor, 7491 Trondheim, Norway

[2] Department of Biostatistics, Epidemiology, Biostatistics and Prevention Institute, University of Zurich, Hirschengraben 84, 8001 Zurich, Switzerland

Received 17 December 2015; revised 4 September 2016; accepted 2 October 2016

The projection of age-stratified cancer incidence and mortality rates is of great interest due to demographic changes, but also therapeutical and diagnostic developments. Bayesian age–period–cohort (APC) models are well suited for the analysis of such data, but are not yet used in routine practice of epidemiologists. Reasons may include that Bayesian APC models have been criticized to produce too wide prediction intervals. Furthermore, the fitting of Bayesian APC models is usually done using Markov chain Monte Carlo (MCMC), which introduces complex convergence concerns and may be subject to additional technical problems. In this paper we address both concerns, developing efficient MCMC-free software for routine use in epidemiological applications. We apply Bayesian APC models to annual lung cancer data for females in five different countries, previously analyzed in the literature. To assess the predictive quality, we omit the observations from the last 10 years and compare the projections with the actual observed data based on the absolute error and the continuous ranked probability score. Further, we assess calibration of the one-step-ahead predictive distributions. In our application, the probabilistic forecasts obtained by the Bayesian APC model are well calibrated and not too wide. A comparison to projections obtained by a generalized Lee–Carter model is also given. The methodology is implemented in the user-friendly R-package BAPC using integrated nested Laplace approximations.

Keywords: Bayesian age–period–cohort model; Cancer projection; Continuous ranked probability score; INLA; Predictive quality.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

Cancer incidence and mortality data are closely monitored in most countries by national registers who collect events by age group, calendar time, gender, and potential other stratification variables. Projections of these data are of strong interest due to demographical changes, but also advances in medical diagnosis and treatment. Based on future trends, financial resources for prevention programs, such as screening procedures, or research programs are distributed (Bray and Møller, 2006). A descriptive analysis on the Lexis diagram (Keiding, 2011) may give hints on the general behavior and allows for qualitative speculations about future time trends. Often also age-standardized trends are inspected and extrapolated (Rosenberg and Anderson, 2011). However, to project data in a more systematic and quantitative manner, model-based approaches are used.

*Corresponding author: e-mail: andrea.riebler@math.ntnu.no, Phone: +47 735 93528, Fax: +47 7359 3524

There exist several statistical models for projections, see Bray and Møller (2006) for an overview. Age–period–cohort (APC) models analyze registry data according to the age group of the individual, the date of the event that is considered (period) and the birth cohort of the individual. The APC model is generally accepted, but its inherent identifiability problem often lead to wrong conclusions, since the observed trend cannot be uniquely assigned to age, period, and cohort (Held and Riebler, 2013; Luo, 2013). Projections based on the APC model are, however, not affected by this problem and hence estimable and interpretable (Holford, 1985). Bayesian APC models are particularly useful to project future cancer burden as they involve no parametric assumptions, see, for example, Berzuini and Clayton (1994), Besag et al. (1995), Knorr-Held and Rainer (2001), Dikshit et al. (2012), Clèries et al. (2013). Bray (2002) provides a comparison of projections derived from linear power models, as well as the classical and Bayesian version of the APC model, and concluded that the Bayesian APC model was the only method to achieve sensible projections.

Surprisingly, APC models are not used in mainstream practice and in particular not for projections (Rosenberg and Anderson, 2011). Reasons may be twofold. First, Bayesian APC models have been criticized for producing too wide credible bands. Clements et al. (2005) have described this problem in an empirical study, but they have aggregated the data from annual to five-year intervals and also have neglected potential overdispersion (OD). Another issue is how to quantify the quality of the forecasts. Møller et al. (2003) assessed point forecasts obtained by the classical APC model, with a couple of modifications. Calibration of probabilistic projections, that is, including the uncertainty of the provided forecasts, has never been investigated systematically and compared to other approaches. A second reason might be a lack of sound and at the same time easy-to-use software. Current software solutions include BAMP (Schmid and Held, 2007), which is a stand-alone software to implement Bayesian APC models using Markov chain Monte Carlo (MCMC) algorithms. Also generic software such as WinBugs (Lunn et al., 2000) or JAGS (Plummer, 2003) might be used, see, for example, Eilstein and Eshai (2012). However, any analysis with MCMC needs to be conducted well to get reliable results. Further, long running times, and the associated convergence checks may push applied scientists away (Bray, 2002; Qui et al., 2010a).

Integrated nested Laplace approximations (INLA) (Rue et al., 2009, 2016) have been shown to be useful to perform MCMC-free inference in APC models (Riebler et al., 2012a) and other Bayesian hierarchical models (Schrödle and Held, 2011; Schrödle et al., 2011). The INLA methodology is available as an R-package, available from www.r-inla.org, which allows the user to apply it to a wide range of different Bayesian hierarchical models. However, the range of options and features is enormous, see Rue et al. (2016), and might be overwhelming if interest lies in one specific model.

In this paper, we present a novel R-package, called BAPC, which builds upon INLA with the aim to forecast future cancer rates and counts within a fully Bayesian inference setting. The applied scientist can easily specify the model and obtains directly the output of interest, such as age-standardized and also age-specific projected rates and observations. To demonstrate the use of the BAPC package for projecting time trends in cancer registry data, we reanalyze yearly data on female lung cancer mortality in five-year age groups from five different countries (Clements et al., 2005). As in Clements et al. (2005) we start predicting the last 10 years for each country using a Bayesian APC model and compare it with the actual realized observations. In contrast to common folklore (Ferlay et al., 2007), we stress that it is not necessary to aggregate the data to five-year periods in an APC analysis and keep the original yearly structure. We further emphasize the crucial importance to use proper scoring rules (Gneiting and Raftery, 2007; Gneiting, 2011) to assess the predictive quality in short and long range. Further, we assess calibration of one-step-ahead probabilistic projections for 25 years with a recently proposed calibration test (Held et al., 2010). A comparison to projections obtained by a generalized Lee–Carter (LC) model shows that the age-specific projections of the APC model provide better scores and are better calibrated.

This paper is organized as follows. Section 2 introduces the data used within this paper. In Section 3 we review the Bayesian APC model and introduce our novel R-package BAPC. In Section 4 we discuss predictive quality assessment based on proper scoring rules and introduce the LC model.

Section 5 assesses retrospective age-stratified and age-specific projections for 10 years. To compare the performance change from short-term to long-term forecasts, we inspect the cumulative average continuous ranked probability score (CRPS) and compare the results to those obtained by a quasi-Poisson version of the LC model. In Section 6 we consider one-step-ahead projections for 25 years and use calibration tests based on the CRPS to compare the predictive quality to the LC model. We summarize our findings in Section 7.

2 Biometrical case study

Our primary data source is the WHO Mortality Database (World Health Organization, 2014). Population estimates and combined cancer mortality counts for the lung, bronchus, and trachea are extracted according to the International Classification of Diseases (ICD). The following codes were used for the different ICD revisions: ICD-7—code A050, ICD-8—code A051, ICD-9—code B101, ICD-10—code 1034 (combined from C33, C340, C341, C342, C343, C348, and C349). We consider the same countries as in Clements et al. (2005), namely Australia (1950–2011), New Zealand (1950–2009), Sweden (1951–2010), the United Kingdom (1950–2010), and the United States of America (1950–2007), but including more recent periods. Data are given by single calendar year and 12 five-year age groups (25–29, 30–34,. . ., 80–84 years).

Since ICD-10 was introduced at different years in the countries of the United Kingdom, mortality counts for the year 2000 are not available. Instead the sum of mortality counts from England and Wales (ICD-9), Northern Ireland (ICD-9) and Scotland (ICD-10) are used as recommended in a notes document provided by the WHO, see http://www.who.int/entity/healthinfo/statistics/notes.zip?ua=1 (updated: July 9, 2012, accessed: 25.04.2014). Australian mortality counts for the year 2005 are also missing in the WHO Mortality database. Here, we use the corresponding counts provided by the Australian Cancer Incidence and Mortality (ACIM) books (http://www.aihw.gov.au/acim-books/). Age-specific mortality rates for all countries are shown in Fig. 1.

3 The APC model

In the absence of OD, the observed incidence or mortality counts y_{ij} in age group i at time point j can be assumed to be Poisson distributed with mean $n_{ij}\lambda_{ij}$, where n_{ij} denotes the corresponding person-time of exposure, assumed to be known. In our application, the age index i runs from 1 to $I = 12$ while the period index j runs from 1 to J with the number of periods J depending on the country considered. The linear predictor $\eta_{ij} = \log(\lambda_{ij})$ is commonly specified as $\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma_k$. Here, μ represents the general level (intercept), and α_i, β_j, γ_k denote age, period, and cohort effects, respectively (Clayton and Schifflers, 1987b). The cohort index k depends on the age group and period index, but also on the width of the age group and period intervals (Holford, 1983, 2006). Here, it is defined as $M \cdot (I - i) + j$, where M indicates that the age group intervals are M times wider than the period interval (Heuer, 1997). In our application M is equal to 5. Of note, this is in contrast to Clements et al. (2005) who aggregated the data artificially to have equal period and age group interval widths of length 5.

3.1 The Bayesian APC model

Bayesian inference treats all unknown parameters as random with appropriate prior distributions. Due to the expectation that effects adjacent in time might be similar, smoothing priors are commonly used for age, period, and cohort effects (Besag et al., 1995; Knorr-Held and Rainer, 2001). A standard choice is the second-order random walk (RW2) (Besag et al., 1995; Rue and Held, 2005), which assumes independent mean-zero normal distributions (with unknown variance) on the second differences of

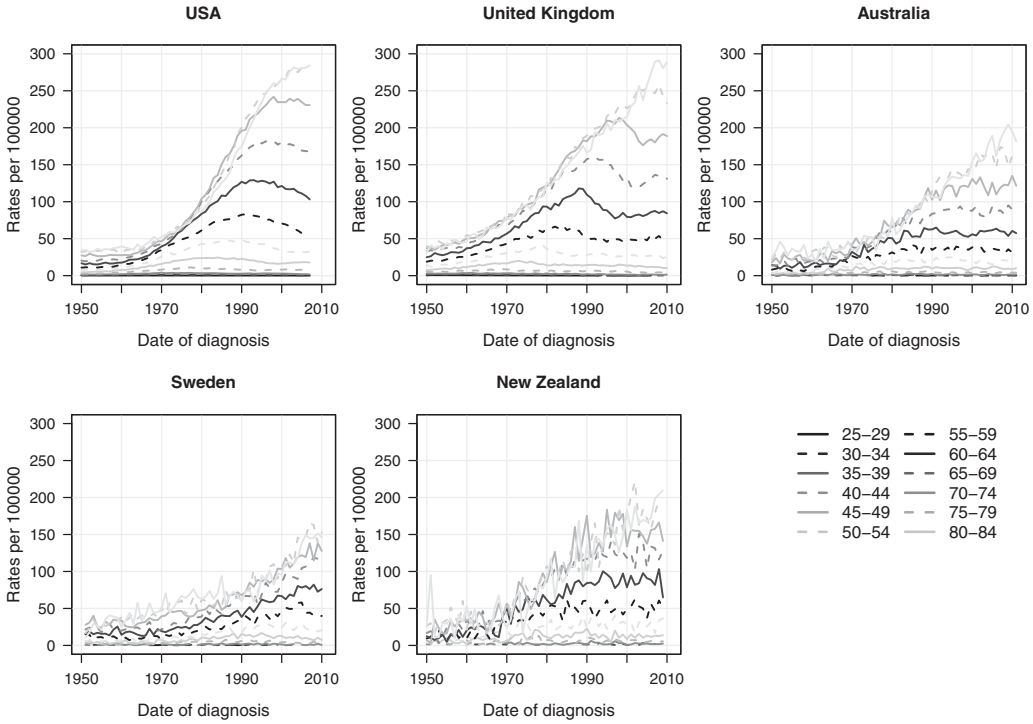

Figure 1 Age-specific lung cancer death rates per 100,000 for females in five different countries.

all time effects. This is a natural target for smoothing, as the second differences in APC models are identifiable (Clayton and Schifflers, 1987b).

Consider the age effects, say, then the RW2 prior is given by

$$f(\boldsymbol{\alpha}|\kappa_\alpha) \propto \kappa_\alpha^{\frac{I-2}{2}} \exp\left(-\frac{\kappa_\alpha}{2}\sum_{i=3}^{I}(\alpha_i - 2\alpha_{i-1} + \alpha_{i-2})^2\right) = \kappa_\alpha^{\frac{I-2}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha}\right)$$

$$Q = \kappa_\alpha \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix},$$

where κ_α^{-1} denotes the variance parameter. Note that Q is not of full rank so the RW2 is an example of an intrinsic Gaussian Markov field (Rue and Held, 2005). The RW2 penalizes deviations from a linear trend and is regarded as the discrete-time analogue of a cubic smoothing spline (Fahrmeir and Tutz, 2001). Alternatively, a linear drift component (Clayton and Schifflers, 1987a) can be specified, which can be seen as a limiting case of the RW2 as the random walk variance goes to zero. To adjust

for OD additional independent mean-zero Gaussian random effects $z_{ij} \sim \mathcal{N}(0, \kappa_z^{-1})$ can be added to the linear predictor $\log(\lambda_{ij})$ (Besag et al., 1995; Knorr-Held and Rainer, 2001).

Depending on the model formulation, that is, the number of RW2 components and the inclusion of an OD component, there are up to four unknown variance parameters. For computational convenience often inverse gamma distributions are assumed for these parameters, where the shape and rate parameter are appropriately defined. Alternatives include half-normal distributions for the standard deviations (Gelman, 2006) and the recently proposed penalized complexity (PC) priors (Simpson et al., 2015).

3.2 Does the identifiability problem affect probabilistic projections?

To ensure the identifiability of the intercept μ the usual sum-to-zero constraints are applied to the set of age, period, and cohort effects. However, there is a second identifiability issue inherent to APC models. In fact, the APC model can be regarded as a partially identified model (Gustafson, 2015, Section 2). Due to the exact linear dependence of age, period, and cohort it is impossible to identify separate contributions of the age, period, and cohort effects (Holford, 2005). There are infinitely many linear transformations that all lead to the same estimated incidence or mortality rate. That means the rate λ_{ij} can be identified, while age, period, and cohort effects cannot. Further, identifiable quantities would be second differences of age, period, or cohort effects. To identify the time effects directly a further constraint is needed in addition to the sum-to-zero constraints (Holford, 1983, 1991). Alternatively, also one set of age, period, or cohort effects could be removed from the model. For more details on partially identified models and their asymptotic behavior we refer to Gustafson (2005, 2015).

Here, we are interested in projections, so that the use of additional constraints can be avoided (Holford, 2005). Assume we are interested in the mortality or incidence rates for the same age groups but t periods ahead into the future, which means

$$\log(\lambda_{iJ+t}) = \mu + \alpha_i + \beta_{J+t} + \gamma_{k+t} + z_{i,J+t},$$

where $k = M \cdot (I - i) + J$. For this purpose, we need to extrapolate the period and cohort effects following the structure of the RW2 model. To be more specific, assume we have data up to period $J \geq 2$, then the period effect at period $J + 1$ will have the conditional distribution

$$\beta_{J+1} \mid \beta_1, \ldots, \beta_J, \kappa_\beta \sim \mathcal{N}\left(2\beta_J - \beta_{J-1}, \kappa_\beta^{-1}\right).$$

If $t > 1$, that is, interest lies not in one-step ahead (as in Section 6) but t-steps ahead forecasts (as in Section 5), this leads to

$$\beta_{J+t} \mid \beta_1, \ldots, \beta_J, \kappa_\beta \sim \mathcal{N}\left((1+t)\beta_h - t\beta_{h-1}, \kappa_\beta^{-1}\left(1 + 2^2 + \cdots + t^2\right)\right).$$

Thus, the conditional mean is given by a linear extrapolation of the last two period effects with cubically increasing variance (Rue and Held, 2005, Section 3.4.1). Analogously, projections for the cohort effects are obtained. It should be noted that not for all age groups a projection of the cohort effects might be needed as the cohort indices run diagonal through the Lexis diagram meaning that certain cohort indices repeat over several period indices (Keiding, 1990). Kuang et al. (2008) show that the identifiability of λ_{iJ+t} depends on the way period and cohort effects are extrapolated. Identifiability is granted using a linear extrapolation scheme, such as the RW2 model. It is not granted using constant extrapolation, as would arise with the alternative random walk of first-order (RW1) model.

3.3 Inference: The BAPC package

We use numerical approximations for Bayesian inference in the APC model. Such an approach works surprisingly well, as already anticipated by Besag et al. (1995, Section 4.4). Specifically, we

use INLA to approximate the posterior marginal distributions directly avoiding any MCMC sampling techniques and therefore also mixing and convergence issues. For methodological details of INLA we refer to Rue et al. (2009). INLA can be implemented via the R-package `INLA`, see www.r-inla.org. To obtain projections in `INLA` pseudo-observations have to be included as missing data, that is, set equal to `NA`. Consequently additional OD effects will be estimated, and period and cohort effects will be extrapolated linearly as described in Section 3.2. Projected linear predictor estimates are also available on request and can be transferred to expected rate projections.

To obtain the predictive distribution for future incidence or mortality counts the Poisson observational noise must be added. This can be done analogously to a MCMC scheme by approximate sampling from the posterior distribution within `INLA` and evaluating the Poisson density at the drawn samples of the linear predictor. To avoid this Monte Carlo approach, we directly compute the mean and variance of the predictive distribution using the law of iterated expectations and the law of total variance, see Supporting Information SA.2 for details.

To make Bayesian APC models with the focus on projections directly available to the demographer or epidemiologist, we present a novel R-package `BAPC` that is available from R-forge (http://r-forge.r-project.org/). The BAPC package is a wrapper of the `INLA` package specific for APC analysis. It facilitates the model specification and offers specialized functions to visualize and extract the output of interest. The goal is to make Bayesian projections of registry data available in routine analysis.

The user can choose whether the period or cohort effects (or both) should be included in the model and also whether an OD component should be incorporated. The only obligatory terms are the age effects. Each time effect can be included in two different ways: as a smooth function based on the intrinsic RW2 model or as a linear drift component. Sum-to-zero constraints are automatically incorporated for RW2 components. In an MCMC context, Besag et al. (1995) discussed that it is not crucial to ensure identifiability of latent parameters as long as the linear predictors $\log(\lambda_{ij})$ are identifiable. By default, `INLA` adds a tiny value ($<1\times10^{-6}$) to the diagonal of the precision matrix of each intrinsic component, which makes prior and posterior proper and thus identifiable. This helps to make numerical computations faster and does affect the posterior distribution of the linear predictors η_{ij} only to a minimal extent. Finally, the user specifies the prior distribution assigned to the unknown variance parameters for the different RW2 and OD components.

The function BAPC provides age-specific projections for the expected rate and number of cases. Further, age-adjusted projections for the rate can be computed when age-specific weights, such as the WHO standard (Ahmad et al., 2001), are provided. Consequently age-specific or age-standardized rates/projections can be plotted using a built-in function. We refer to Supporting Information SA.1, which provides R-code to illustrate the usage of the BAPC package in practice using the data for United States analyzed here. As second example we reanalyze the data presented in Holford (1983), Besag et al. (1995), where age groups and periods are given for the same interval length. In the latter example, we further illustrate how the prior distributions for the variance parameters of the random effects can be changed. In the paper we used version 0.0-1462622767 of the R-package `INLA` and version 0.0.33 of `BAPC`.

4 Predictive quality assessment

In the following two sections, we will assess the predictive quality of the Bayesian APC model based on both retrospective and one-step-ahead projections. As a first check of calibration of predictions, the empirical coverage of the 50%, 80%, and 95% predictive credible bands can be computed. However, this is only recommended for one-step-ahead projections as otherwise the dependency structure between the projections needs to be addressed. To assess the quality of projections systematically, we use methodology based on proper scoring rules (Gneiting and Raftery, 2007; Gneiting and Katzfuss,

2014). The results are compared to the commonly used generalized LC model (Lee and Carter, 1992; Renshaw and Haberman, 2006).

4.1 Proper scoring rules and calibration tests

Proper scoring rules assess both calibration and sharpness of probabilistic predictions jointly. Here, we use the CRPS, which is closely related to the Brier score (Spiegelhalter, 1986) and commonly used in meteorological and economical applications (Gneiting and Raftery, 2007; Gneiting and Katzfuss, 2014). Specifically, let y_{ij} denote the actually observed number of cases in age group i at period j, and μ_{ij} and σ_{ij} the mean and standard deviation of the respective (normally distributed) predictive distribution, which are obtained as shown in Supporting Information SA.2. The CRPS for the ij-th normal prediction can be computed as

$$\mathrm{CRPS}_{ij} = \sigma_{ij}\left[\tilde{y}_{ij}\{2\,\Phi(\tilde{y}_{ij}) - 1\} + 2\,\phi(\tilde{y}_{ij}) - 1/\sqrt{\pi}\right],$$

where $\tilde{y}_{ij} = (y_{ij} - \mu_{ij})/\sigma_{ij}$ are the standardized observed number of cases with respect to their predictive distribution, while $\phi(\cdot)$ and $\Phi(\cdot)$ denote the density and distribution functions, respectively, of the standard normal distribution.

If $\sigma_{ij} = 0$, the CRPS reduces to the absolute error (AE), which can be easily seen from the equivalent formulation

$$\mathrm{CRPS}_{ij} = \mathrm{E}|Y_{ij} - y_{ij}| - \frac{1}{2}\,\mathrm{E}|Y_{ij} - Y'_{ij}|,$$

here Y_{ij} and Y'_{ij} are independent realizations from the predictive (normal) distribution (Gneiting and Raftery, 2007). The AE is commonly used to assess the quality of point predictions. Sometimes the absolute value of the difference between observed and predicted numbers of cases relative to the observed number of cases has been used instead (Møller et al., 2003; Lee et al., 2011). However, this tends to support severe underforecasts and its general suitability as a scoring function has been recently questioned (Gneiting, 2011).

Both AE and CRPS are negatively oriented, that is, the smaller the better the forecast. In addition, both can be reported in the same unit as the observations (Gneiting and Raftery, 2007), here in the number of cancer cases. In contrast to the AE, the CRPS is not only based on the point projection μ_{ij}, but also on the predictive standard deviation σ_{ij}.

The cumulative average $\overline{\mathrm{CRPS}}_j$ denotes the mean CRPS across age group and all periods up to and including period j beginning at the first period to be predicted. It can be used to compare the performance change from short-term to long-term forecasts (Riebler et al., 2012a). The mean CRPS, denoted as $\overline{\mathrm{CRPS}} = \overline{\mathrm{CRPS}}_J$, can be used as an overall criterion to compare the quality of different probabilistic prediction models (Gneiting and Raftery 2007). The cumulative average $\overline{\mathrm{AE}}_j$ and mean AE, that is, $\overline{\mathrm{AE}}$, are defined analogously.

To assess calibration further, we use a recently proposed calibration test based on $\overline{\mathrm{CRPS}}$ (Held et al., 2010). Consider the test statistic

$$z = \frac{\overline{\mathrm{CRPS}} - \mathrm{E}_0(\overline{\mathrm{CRPS}})}{\sqrt{\mathrm{Var}_0(\overline{\mathrm{CRPS}})}}, \tag{1}$$

where E_0 and Var_0 denote the mean and variance of the CRPS under perfect calibration, where the data-generating distribution is the same as the forecast distribution, see Held et al. (2010). When calculating $\mathrm{Var}_0(\overline{\mathrm{CRPS}})$ we assume independence of the components CRPS_{ij} of $\overline{\mathrm{CRPS}}$, which does hold for one-step-ahead predictions (Seillier-Moiseiwitsch et al., 1992; Seillier-Moiseiwitsch and Dawid, 1993). We assume that predictions in different age groups are also (approximately) independent, so the test statistic

z is approximately standard normal distributed and a (two-sided) p-value can be easily computed. The smaller the p-value, the larger the evidence that the forecasts are miscalibrated. Furthermore, the sign of the test statistic z indicates, whether the predictions are over- (negative sign) or underdispersed (positive sign) relative to the observations. In the first case, the empirical coverage of the prediction intervals tends to be larger than the nominal coverage, in the second case the empirical coverage tends to be smaller. If the predictions are "too wide," then they are overdispersed and so the sign of the test statistic will be negative.

4.2 The LC model

We compare the projections and their calibration obtained by the Bayesian APC model to those obtained by a generalized version of the LC model (Lee and Carter, 1992; Renshaw and Haberman, 2006) with Poisson error structure where

$$\log(\lambda_{ij}) = \alpha_i + \beta_i \cdot \kappa_j. \tag{2}$$

Here, α_i represents the age effect, κ_j measures the mortality trend across periods, and β_i denotes the age-specific deviation of mortality change from the general trend. It is further possible to include a second bilinear term related to a cohort effect, so that

$$\log(\lambda_{ij}) = \alpha_i + \beta_i^{(0)} \iota_{j-i} + \beta_i^{(1)} \cdot \kappa_j, \tag{3}$$

however, here it is implicitly assumed that the width of age group and period intervals is the same. Further, it is recommended to set $\beta_i^{(0)} = 1$ in order to resolve forecasting issues (Haberman and Renshaw, 2011; Booth et al., 2013). All presented models can be implemented using the R-packages `demography` (Hyndman, 2014) and `ilc` (Butt and Haberman, 2010). We note that cells with zero observations are ignored in the maximum-likelihood (ML) fitting procedure of `ilc` and `demography`. Here, we present the results of the most commonly used model, the generalized LC model, as presented in (2). Further, we tried the extended version (3) including a cohort effect where we set $\beta_i^{(0)}$ to 1, however here the ML procedure returned errors indicating nonconvergent deviance. The mean and standard deviation of the predictive distribution are derived as outlined in Supporting Information SA.3 to compute PIT histograms (Gneiting et al., 2007) and the test statistic (1). The predictive distribution of age-standardized projections is not directly available since the covariance structure between the linear predictor terms is not returned. Thus, only point forecasts can be derived.

5 Retrospective projections for the last 10 years

Following Clements et al. (2005) projections were made for each of the five countries by omitting the observed number of mortality cases for all age groups in the respective last 10 years, corresponding to $10 \cdot 12 = 120$ age-specific projections per country. Clements et al. (2005) criticized the wide credible bands provided by the Bayesian APC model. However, they required to have the same age group and period interval width and aggregated the annual periods to five-year intervals, while keeping the annual data structure when fitting the frequentist generalised additive models. This may have led to biased performance assessment using plug-in and predictive deviance as the predictive distribution of the Bayesian APC model might be wider. Here, we keep the yearly data structure and fit models with and without additional adjustments for OD (Knorr-Held and Rainer, 2001). We compare projected mortality counts and rates to the realized observations by means of age-specific plots. Further we use proper scoring rules to assess how the predictive quality changes from short-term to long-term projections. A comparison to the results of the generalized LC model is given. Age-standardized projected rates are computed based on the world standard population and compared to the aggregated observed values (Ahmad et al., 2001).

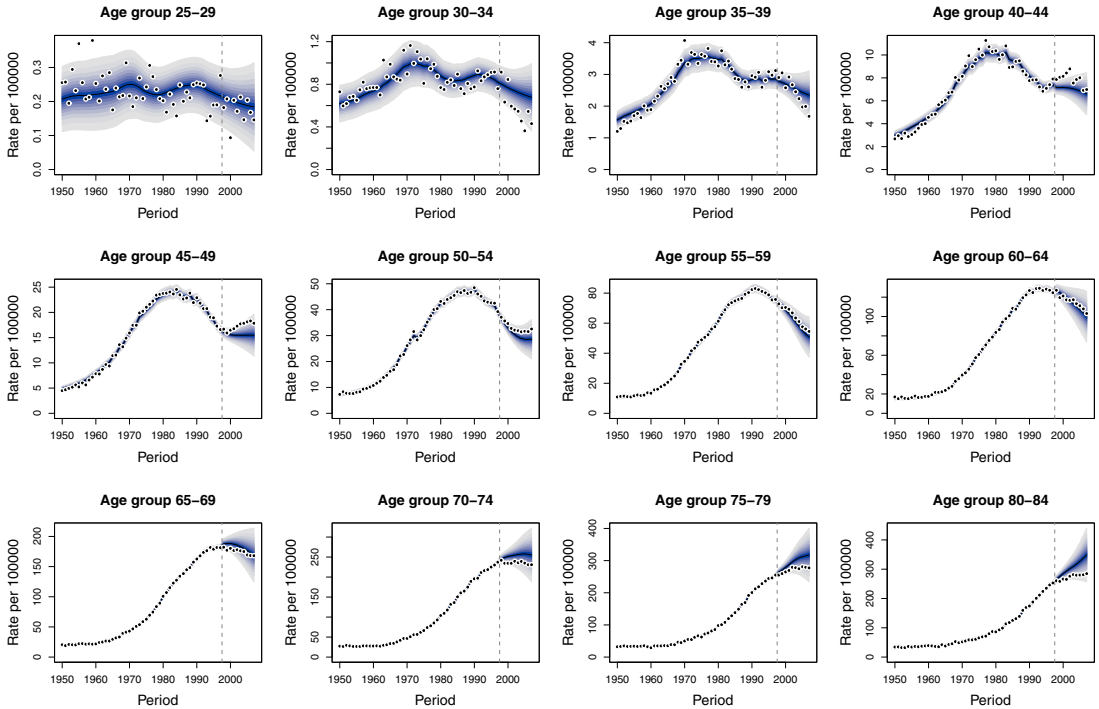

Figure 2 Female lung cancer mortality forecasts in the United States. The fan shows the predictive distribution between the 5% and 95% quantile, whereby the shaded bands show prediction intervals in increments of 10%. The predictive mean is shown as solid line. Observed number of cases are shown as a filled circle. The vertical dashed line indicates where prediction started.

Figure 2 shows observed and predicted numbers for female lung cancer mortality in the United States for all age groups. The different shadings indicate pointwise credible intervals (Spiegelhalter et al., 2011). The central interval represents 10% probability, and the largest interval 90% probability. The corresponding figures for the other countries are shown in the Supporting Information SA.4. The plots show that the projections seem reasonable for almost all age groups and countries.

To compare the performance change from short-term to long-term forecasts, Fig. 3 shows the cumulative average $\overline{\text{CRPS}}_j$ (solid lines). The curves of the Bayesian APC model (black) are always below those of the LC model (gray) indicating better predictive quality. For Australia, Sweden, and New Zealand the scores for the Bayesian APC model stay fairly constant while a decrease in projection quality over time is seen for the larger countries United States and United Kingdom. However, the cumulative average $\overline{\text{CRPS}}_j$ of the LC model increases faster than for the APC model in all countries, which indicates lower projection quality with increasing time compared to the APC model. For comparison Fig. 3 also shows the curves for the easier-to-interpret cumulative average $\overline{\text{AE}}_j$ (dotted), which lie over those for the CRPS.

Figure 4 shows projected age-standardized rates for all countries together with the observed rates. For age standardization the WHO world population was used as the standard population (Ahmad et al., 2001). The weights are scaled so that they sum to 1. Age-standardized rates per 100,000 inhabitants were computed as weighted sum of the corresponding age-specific rates. The pointwise credible bands get clearly wider when going more away from the fitted data and may be regarded as too wide when looking at the United States. For the United Kingdom, in contrast, the width seems to be reasonable. To analyze the quality of the projections further, we will investigate in Section 6

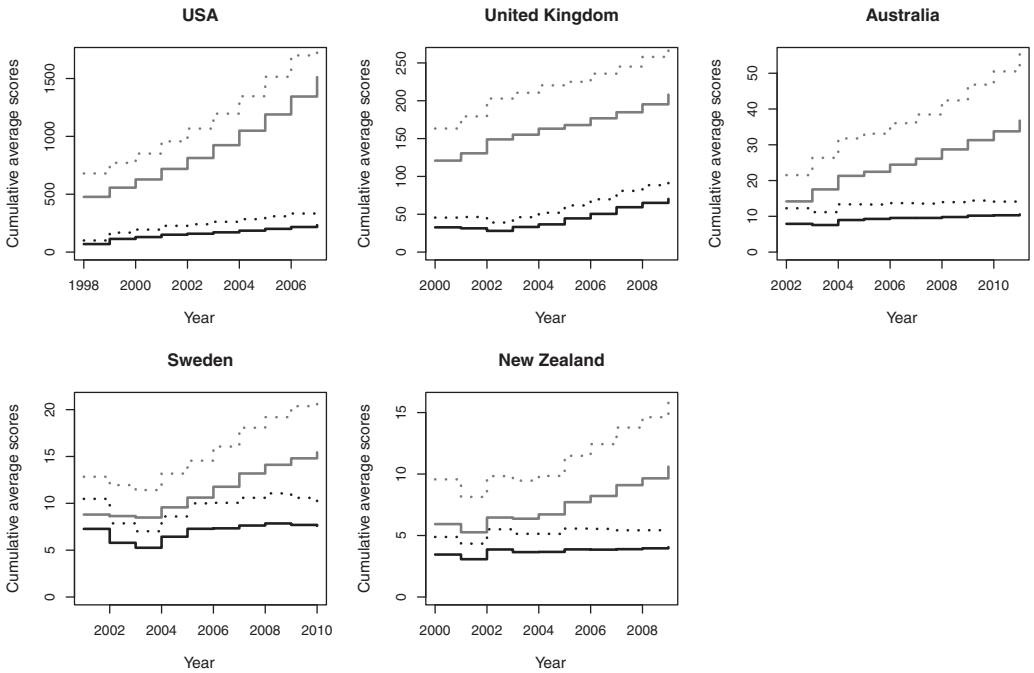

Figure 3 Cumulative average of mean absolute errors (dotted) and continuous ranked probability scores (line) across age groups for all five countries obtained by the Bayesian APC model (black) and the Lee–Carter model (gray). Note the values at the last timepoints correspond to the overall $\overline{\text{AE}}$ and $\overline{\text{CRPS}}$ scores averaged over all 120 projections, respectively.

the calibration and sharpness of age-specific and age-standardized forecasts based on consecutive one-step-ahead predictions. The results will again be compared to the generalized LC model.

6 One-step-ahead projections for 25 years

In this section, we project for each country the age-specific lung cancer mortality counts for the most recent 25 years based on a one-step-ahead procedure. That means, that for projecting the mortality counts for all age groups in the year 2000, say, we use all data up to 1999, while for projecting the observations for 2001 we use all data up to 2000, and so on. In total this leads to $25 \cdot 12 = 300$ age-specific projections for each country. Those can be ultimately aggregated to 25 age-standardized projections, see Supporting Information SA.2.

6.1 Age-specific projections

The empirical coverage shown in Table 1 indicates that the projections for United States and the United Kingdom are underdispersed for both the APC model and the LC model. However, the empirical coverage of the APC model including OD is closer to the nominal level than that of the LC model. For the other three countries, the empirical coverage for the APC model, with or without adjustment for OD, is very close to the nominal level with a difference never larger than 4 percentage points. The LC model has larger disagreement, in particular for Sweden. This is reflected in a very small p-value of the CRPS calibration test ($p = 0.0001$), whereas there is no evidence for

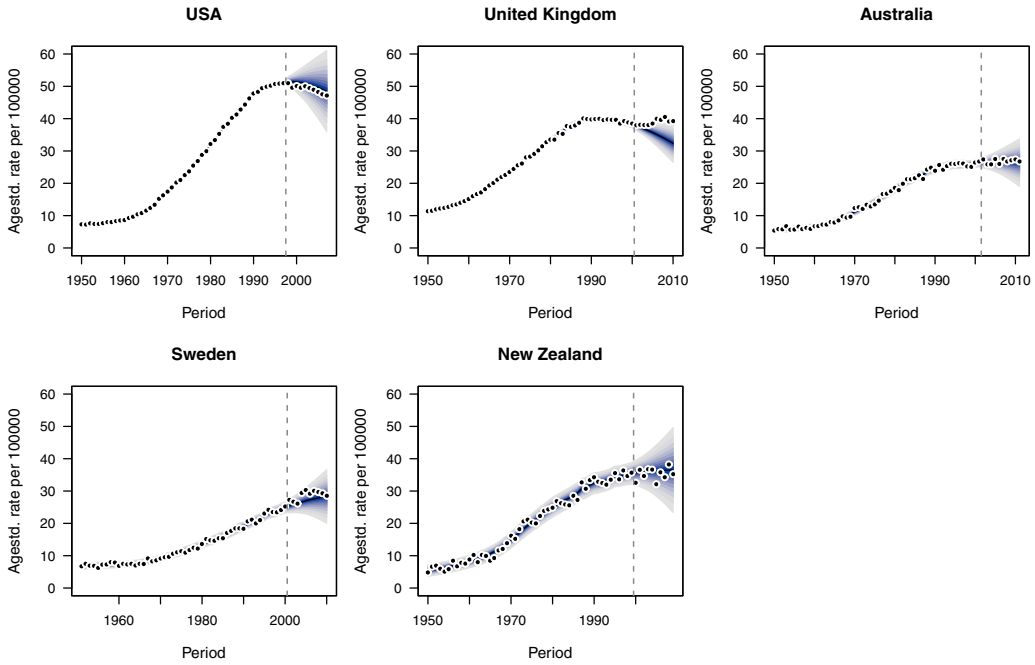

Figure 4 Retrospective projections of lung cancer mortality rates by country for females 25–84 years. Shown are observed rates (dots) together with the predictive distribution between the 5% and 95% quantile, whereby the shaded bands show prediction intervals in increments of 10%. The predictive mean is shown as solid line and the vertical dashed line indicates where prediction started.

Table 1 Empirical coverage of the one-step-ahead predictive credible bands for three different credible levels 50%, 80%, and 95%.

	APC			APC (no overdis.)			Lee–Carter		
Credible level	50%	80%	95%	50%	80%	95%	50%	80%	95%
United States	36%	66%	92%	36%	56%	76%	37%	57%	86%
United Kingdom	44%	67%	86%	35%	61%	81%	38%	64%	82%
Australia	51%	83%	97%	46%	78%	94%	45%	78%	96%
Sweden	49%	82%	95%	46%	79%	93%	41%	74%	91%
New Zealand	52%	82%	94%	49%	80%	93%	56%	84%	96%

Notes: Shown are results obtained with the Bayesian APC model, with and without the inclusion of overdispersion parameters, and the generalized Lee–Carter model for all countries.

miscalibration of the APC model after adjusting for OD ($p = 0.63$), see Table 2. However, both for United States and the United Kingdom, there is strong evidence that the projections are miscalibrated for both APC models as well as the LC model ($p < 0.0001$). Interestingly, all z-statistics are positive for these two countries, indicating underdispersed projections. This corresponds to too low empirical coverage as shown in Table 1. Including OD parameters for Australia, there is weak evidence that the APC projections are overdispersed ($z = -1.941$, $p = 0.053$) and that the LC projections are underdispersed ($z = 1.761$, $p = 0.078$). For New Zealand, there is moderate evidence that the LC projections

Table 2 Mean absolute error $\overline{\text{AE}}$, mean predictive standard deviation $\overline{\text{SD}}$ and mean continuous ranked probability score $\overline{\text{CRPS}}$ with z-statistic and p-value from the corresponding calibration test.

Model	Country	$\overline{\text{AE}}$	$\overline{\text{SD}}$	$\overline{\text{CRPS}}$	z	p-Value
APC	United States	155.59	149.76	108.40	5.25	< 0.0001
	United Kingdom	39.84	41.16	28.70	4.39	< 0.0001
	Australia	10.64	14.67	7.47	−1.94	0.053
	Sweden	7.95	10.13	5.58	−0.48	0.63
	New Zealand	4.86	6.38	3.49	−0.66	0.51
APC (no overdis.)	United States	115.58	83.40	88.20	17.20	< 0.0001
	United Kingdom	37.75	31.18	27.68	11.16	< 0.0001
	Australia	10.70	12.45	7.52	1.44	0.15
	Sweden	7.93	9.01	5.57	1.94	0.052
	New Zealand	4.87	5.94	3.49	0.84	0.40
Lee–Carter	United States	594.15	508.12	415.93	8.79	< 0.0001
	United Kingdom	141.56	100.90	101.78	15.11	< 0.0001
	Australia	21.51	23.91	14.72	1.76	0.078
	Sweden	11.33	11.99	8.06	3.88	0.0001
	New Zealand	5.96	8.70	4.27	−2.56	0.01

Notes: Shown are the results for the one-step-ahead projections obtained with the Bayesian APC model, with and without the inclusion of overdispersion parameters, and the generalized Lee–Carter model for all countries.

are overdispersed ($z = -2.560$, $p = 0.01$), whereas there is no evidence that the APC projections are miscalibrated.

Importantly, the mean CRPS scores provided in Table 2 clearly show that the APC model always provides better projections than the LC model with substantially lower mean scores. For example, for United States the mean CRPS is 108.4 cases for the APC model including OD, but 415.9 cases for the LC model, that is, a nearly fourfold increase. For smaller countries the scores and the score differences are smaller, for example, 5.6 cases (APC) versus 8.1 cases (LC) for Sweden, but the order is always the same with better projections based on the APC model. In addition to the CRPS also the AE is given. Of note the ratio of $\overline{\text{AE}}$ of the Bayesian APC model including OD parameters versus the $\overline{\text{AE}}$ of the LC model is roughly the same as the ratio of the corresponding $\overline{\text{CRPS}}$ scores.

To summarize, the APC model produces substantially better one-step-ahead forecasts than the LC model. There is evidence for miscalibration of the forecasts for United States and the United Kingdom, but the forecast distributions are here too narrow, not too wide. Only for Australia there is weak evidence of overdispersed APC forecasts, but the empirical coverage shown in Table 1 is only slightly above the nominal level (83% and 97% on nominal 80% and 95% level). Removing the OD parameters in the analysis of the United States and United Kingdom mortality leads to improved one-step-ahead predictions with lower mean CRPS scores, see Table 2 (88.20 and 27.68 cases, respectively). This is caused by smaller AEs and increased sharpness, that is, increased concentration of the predictive distributions indicated by the lower mean predictive standard deviation $\overline{\text{SD}}$ (Gneiting and Katzfuss, 2014). However, the corresponding z-scores of the calibration test are even larger now (17.20 and 11.16, respectively) reflecting also increased miscalibration due to underdispersed predictions. This can also be seen in the corresponding empirical coverage, see Table 1, which is worse than when accounting for OD. It should be noted that adjusting for OD is known to influence the sharpness of probabilistic predictions, but it is slightly surprising that here also the point forecasts changed as indicated by lower $\overline{\text{AE}}$ scores.

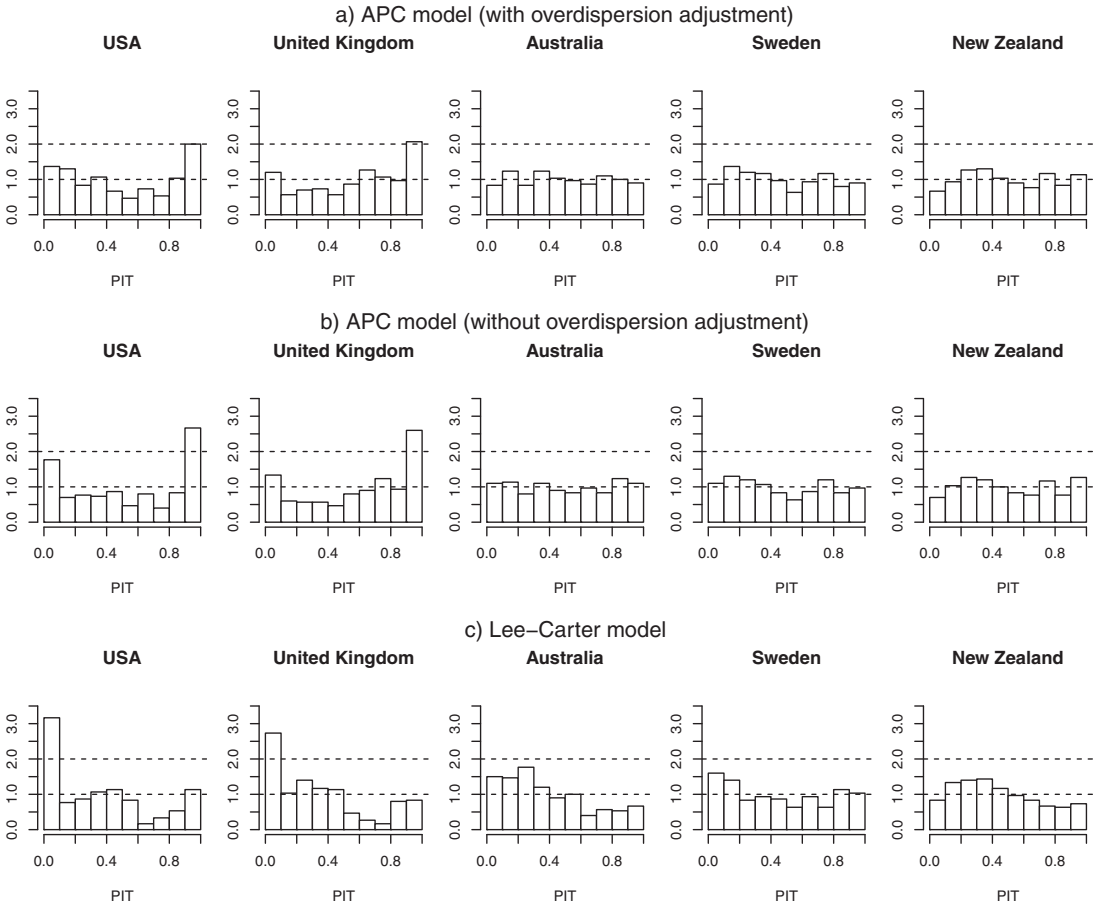

Figure 5 PIT histograms. Shown are the results obtained with the Bayesian APC model, with (a) and without (b) adjustment for overdispersion, and the generalized Lee–Carter model (c) for all countries.

Figure 5 shows the PIT histograms based on all projections made for each of the countries, which should ideally follow a uniform distribution (Gneiting and Raftery, 2007). In the top row the results of the APC model including OD parameters are shown. It can be seen that the PIT histogram for Australia looks almost perfect, while the ones for United States and United Kingdom deviate slightly from uniformity. Removing the OD parameters (middle row) the PIT histograms for United States and United Kingdom get a stronger U-shape indicating too narrow predictive distributions. However, as shown in Table 1, adjusting for OD for these countries leads too worse mean scores. The increased AE, in particular for the United States indicates worse point forecasts. The calibration on the other hand is improved as indicated by the lower z-scores, whereby the p-values still indicate miscalibration. For the other countries we see no strong differences. The PIT histograms for the LC model deviate all moderately from a uniform distribution. For New Zealand a slight bump shape with lower values close to 0 and 1 can be seen, which indicates a too wide predictive distribution and thus OD. In contrast, for Sweden we note a slight U-shape indicating a too narrow predictive distribution and thus underdispersion. Both observations are supported by Table 1.

Table 3 $\overline{\text{AE}}$, and $\overline{\text{CRPS}}$ with corresponding z-statistic and p-value used in the calibration test.

	$\overline{\text{AE}}$	$\overline{\text{CRPS}}$	z	p-Value
United States	511.29	357.45	−1.38	0.17
United Kingdom	127.76	86.97	−0.50	0.62
Australia	49.43	33.89	−0.47	0.64
Sweden	33.02	21.31	−0.06	0.95
New Zealand	19.73	13.82	−0.80	0.42

Notes: Shown are the results for the age-standardized projections obtained with the Bayesian APC model (including overdispersion parameters) for all countries.

6.2 Age-standardized projections

We also inspected age-standardized projections for the Bayesian APC model including OD parameters, where we used again the WHO world population as reference population (Ahmad et al., 2001). Due to the aggregation the number of projections reduces drastically from 300 to 25, which makes the assessment of predictive quality more difficult. Table 3 shows mean AE and CRPS scores for age-standardized projections for the Bayesian APC model. The results obtained by the corresponding calibration test for the CRPS are given as well. The mean CRPS is increased compared to the age-specific projections. However, there is no evidence that the projections are miscalibrated. The z-scores are negatively oriented but the absolute values are not suspicious (< 1.4) so that the calibration test shows no evidence for miscalibration. Of note, the CRPS cannot be computed for the LC models, as the predictive distribution for age-standardized projection is not available. However, the mean AE of the Bayesian APC model was never more than 15% of the mean AE of the LC model for all countries considered.

6.3 Prior sensitivity

In this analysis we used inverse gamma priors with shape parameter $a = 1$ and rate parameter $b = 0.00005$ for the RW2 variance parameters of the time effects (age, period, and cohort), and an inverse gamma prior with $a = 1$ and $b = 0.005$ for the OD variance (Knorr-Held and Rainer, 2001). To assess sensitivity we matched half-normal priors (for the respective standard deviations), see Roos et al. (2015) for details, to these choices and also used PC priors (Simpson et al., 2015). We found that results both for age-specific and age-standardized projections are almost not prior sensitive and that the same conclusions as shown for the gamma priors can be drawn.

To check the effect when changing the rate parameter of the inverse gamma priors, we compared the results to those obtained when using the hyperparameters proposed by Smith and Wakefield (2016) for either only the RW2 variance parameters, only the OD variance or all variances. The detailed results can be found in the Supporting Information SA.7. Also here the results seem fairly stable and not very sensitive to the prior change.

7 Discussion

Bayesian APC models are very popular to analyze and project age-specific cancer incidence or mortality data. However, they have not yet found the way into the routine analysis of epidemiologists. For practice use often simple models are favored (Dyba and Hakulinen, 2008) and thus Bayesian approaches that require potential long MCMC runs to reach convergence are not attractive (Bray, 2002; Qiu et al., 2010a). Here, we presented the new R-package `BAPC` that uses INLA to perform fully Bayesian inference

and thus avoids MCMC sampling completely. Model specification is straightforward and summary plots and estimates are directly available, see Supporting Information SA.1.

Another point of criticism is that Bayesian APC models would produce too wide prediction intervals (Clements et al., 2005). Some authors question whether prediction intervals should be reported at all and whether such intervals are interpretable (Møller et al., 2005; Dyba and Hakulinen, 2008; Rutherford et al., 2012). For example, the software package Nordpred developed by the Cancer Registry of Norway (http://www.kreftregisteret.no/software/nordpred) only reports point forecasts. However, as there will always be substantial uncertainty in demographic forecasts we think that forecasts should be probabilistic (Gneiting and Katzfuss, 2014).

Clements et al. (2005) analyzed lung cancer mortality of females in five different countries that vary by population size, namely United States, United Kingdom, Australia, Sweden, New Zealand, and criticized the width of prediction intervals obtained by Bayesian APC models as being too wide. In contrast, the prediction intervals of a two-dimensional generalised additive models were found to be narrower. However, having data for more recent years today the question arises whether these intervals were in fact not too optimistic. Furthermore, aggregation of annual to five-year data and omission of OD when implementing the Bayesian APC model may have contributed to this finding.

In this paper we reanalyzed the data of Clements et al. (2005) (updated to the current years) and systematically investigated the predictive quality of Bayesian APC models. Based on retrospective and one-step-ahead projections, we found that the Bayesian APC model provides sensible forecast. The projection quality stayed for three of the five countries almost constant when going from short-term to long-term projections, while a slight decrease was seen for the larger countries United States and United Kingdom. We compared these results to the quasi-Poisson version of the LC model that showed worse performance. Using a calibration test based on the CRPS for one-step-ahead projections we further found no evidence that the projection intervals for neither age-specific nor age-standardized projections are too wide. Further, the Bayesian APC model was always superior compared to the generalized LC model. In fact, the LC model seems to suffer from the bilinear term, which implicitly assumes that age groups who have seen the greatest mortality improvements in the past will also see the greatest improvements in the future (Alai and Sherris, 2014). In contrast to the APC model the original LC formulation does not include cohort effects, however Alai and Sherris (2014) showed that the bilinear effect between age and period is actually comparable to the main cohort effect in the APC model.

Knorr-Held and Rainer (2001) discussed the incorporation of covariate effects in the Bayesian APC model to improve the predictive quality for lung cancer mortality. They replaced thereby the period parameters with a regression variable related to the number of cigarettes sold and to the average tar content per cigarette. Such a replacement is linked to certain assumptions, such as that other relevant factors, which could be attributed to the period effects, are time constant. The question whether those assumptions are justifiable, strongly depends on the application. Further, besides from finding precise covariate data reflecting changes in exposure, say, it might be hard to determine the corresponding time lag with which this covariate reflects in changes of cancer mortality. The larger this time lag the harder it might then be to obtain this covariate information for early periods (Knorr-Held and Rainer, 2001). Currently the incorporation of covariate effects is not supported in `BAPC`, but we consider appropriate extensions in the future.

For predicting further into the future than 10 years it might be sensible to level off the exponential trend assumed by the Bayesian APC model presented here. Possibilities include using a power link instead of a log link or reducing the impact of current trends (Møller et al., 2002). Recent work of Jürgens et al. (2014) promotes to use the power link where the power is not fixed, as in Nordpred, but assumed to be random. Furthermore, the use of a structured interaction effect between age and period might be considered (Havulinna, 2014).

We think that the Bayesian APC model is a well-suited model to project cancer incidence or mortality and that its usage is now greatly simplified using our new R-package `BAPC`. No advanced coding is required on the user-side, no convergence checks are necessary, and plots and results of interest are

directly available. The specification of prior distributions for the variance parameters is of course still needed (Qiu et al., 2010b), where the recent development of PC priors may be useful.

In classical analyses, zero counts in the observed data may lead to problems in model fitting (Baker and Bray, 2005). In the data sets considered zero counts were mostly observed for Australia, Sweden, and New Zealand, while no sparsity was present for the United States. In Supporting Information SA.6 we investigate the effect of excluding data from the youngest age group on the predictive quality. We conclude that for the smaller countries where sparsity is present there is almost no effect on one-step ahead predictions while inclusion of all data seems beneficial when prediction time increases. For the United States exclusion of data from the youngest age group seems to slightly improve both short- and long-term predictions. Stronger varying cohort effects, see Supporting Information SA.5, might be a reason.

Following Clements et al. (2005), we analyzed the data from the five different countries given in this paper separately. However, when analyzing data of countries that are related or data for related causes of mortality, a joint analysis using multivariate APC models may be beneficial, see Riebler and Held (2010). Riebler et al. (2012a) borrowed strength from related populations to improve forecasting performance. In future work we consider extending the `BAPC` package to handle these settings as well.

It is important to note that this package is also interesting outside cancer surveillance. For example, Carreras and Gorini (2013) analyzed the prevalence of former smokers in Italy and used Bayesian APC models to project future trends, while Riebler et al. (2012b) used Bayesian APC models to analyze and predict suicide mortality. So far, the `BAPC` package assumes count data, but we plan to generalize this in future releases.

Acknowledgments The authors like to thank the Editor, Associate Editor, and three anonymous reviewers for their helpful comments and suggestions, which lead to an improved version of the paper. Further, we like to thank Håvard Rue for useful discussions.

Conflict of interest
The authors have declared no conflict of interest.

References

Ahmad, O., Boschi-Pinto, C., Lopez, A. D., Murray, C. J. L., Lozano, R., and Inoue, M. (2001). Age standardization of rates: a new WHO standard, in *GPE Discussion Paper Series* (GPE Discussion Paper No. 31). The World Health Organization, Geneva.

Alai, D. H. and Sherris, M. (2014). Rethinking age-period-cohort mortality trend models. *Scandinavian Actuarial Journal* **2014**, 208–227.

Baker, A. and Bray, I. (2005). Bayesian projections: what are the effects of excluding data from younger age groups? *American Journal of Epidemiology* **162**, 798–805.

Berzuini, C. and Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in Medicine* **13**, 823–838.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Booth, H., Hyndman, R. J., and Tickle, L. (2013). *Prospective life tables*, Computational Actuarial Science with R. CRC Press, Boca Raton, FL, pp. 319–344.

Bray, F. and Møller, B. (2006). Predicting the future burden of cancer. *Nature Reviews Cancer* **6**, 63–74.

Bray, I. (2002). Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Journal of the Royal Statistical Society—Series C* **51**, 151–164.

Butt, Z. and Haberman, S. (2010). *Ilc: A Collection of R Functions for Fitting a Class of Lee-Carter Mortality Models using Iterative Fitting Algorithms* Actuarial Research Paper No. 190, Cass Business School.

Carreras, G. and Gorini, G. (2013). Time trends of Italian former smokers 1980–2009 and 2010–2030 projections using a Bayesian age period cohort model. *International Journal of Environmental Research and Public Health*, **11**, 1–12.

Clayton, D. and Schifflers, E. (1987a). Models for temporal variation in cancer rates. I: age-period and age-cohort models. *Statistics in Medicine* **6**, 449–467.

Clayton, D. and Schifflers, E. (1987b). Models for temporal variation in cancer rates. II: age-period-cohort models. *Statistics in Medicine* **6**, 469–481.

Clements, M., Armstrong, B., and Moolgavkar, S. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics* **6** , 576–589.

Clèries, R., Martínez, J. M., Moreno, V., Yasui, Y., Ribes, J., and Borràs, J. M. (2013). Predicting the change in breast cancer deaths in Spain by 2019: a Bayesian approach. *Epidemiology* **24**, 454–460.

Dikshit, R. P., Yeole, B., Nagrani, R., Dhillon, P., Badwe, R., and Bray, F. (2012). Increase in breast cancer incidence among older women in Mumbai: 30-year trends and predictions to 2025. *Cancer Epidemiology* **36**, e215–e220.

Dyba, T. and Hakulinen, T. (2008). Do cancer predictions work? *European Journal of Cancer* **44**, 448–453.

Eilstein, D. and Eshai, K. (2012). Lung and breast cancer mortality among women in France: future trends. *Cancer Epidemiology* **36**, e341–e348.

Fahrmeir, L. and Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2nd edn.). Springer, Berlin, DE.

Ferlay, J., Autier, P., Boniol, M., Heanue, M., Colombet, M., and Boyle, P. (2007). Estimates of the cancer incidence and mortality in Europe in 2006. *Annals of Oncology* **18**, 581–592.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (with discussion). *Bayesian Analysis* **1**, 515–534.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**, 746–762.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B (Methodological)* **69**, 243–268.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application* **1**, 125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Gustafson, P. (2005). On model expansion, model contraction, identifiability and prior information: two illustrative scenarios involving mismeasured variables. *Statistical Science* **20**, 111–140.

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*. Chapman & Hall/CRC, Boca Raton, FL.

Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics* **48**, 35–55.

Havulinna, A. S. (2014). Bayesian age–period–cohort models with versatile interactions and long-term predictions: mortality and population in Finland 1878–2050. *Statistics in Medicine* **33**, 845–856.

Held, L. and Riebler, A. (2013). Comment on Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort (APC) problem. *Demography* **50**, 1977–1979.

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics* **66**, 1295–1305.

Heuer, C. (1997). Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics* **53**, 161–177.

Holford, T. (2005). Age-period-cohort analysis, In: Armitage, P., and Colton, T.(Eds.), *Encyclopaedia of Biostatistics* (2nd edn.). John Wiley and Sons, West Sussex, pp. 105–123.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics* **39**, 311–324.

Holford, T. R. (1985). An alternative approach to statistical age-period-cohort analysis. *Journal of Chronic Diseases* **38**, 831–836.

Holford, T. R. (1991). Understanding the effects of age, period and cohort on incidence and mortality rates. *Annual Reviews of Public Health* **12**, 425–457.

Holford, T. R. (2006). Approaches to fitting age-period-cohort models with unequal intervals. *Statistics in Medicine* **25**, 977–993.

Hyndman, R. J. (2014). *Demography: Forecasting Mortality, Fertility, Migration and Population Data*. URL http://CRAN.R-project.org/package=demography, with contributions from Heather Booth and Leonie Tickle and John Maindonald. R package version 1.17.

Jürgens, V., Ess, S., Cerny, T., and Vounatsou, P. (2014). A Bayesian generalized age–period–cohort power model for cancer projections. *Statistics in Medicine* **33**, 4627–4636.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **332**, 487–509.

Keiding, N. (2011). Age-period-cohort analysis in the 1870s: diagrams, stereograms, and the basic differential equation. *Canadian Journal of Statistics* **39**, 405–420.

Knorr-Held, L. and Rainer, E. (2001). Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics* **2**, 109–129.

Kuang, D., Nielsen, B., and Nielsen, J. P. (2008). Forecasting with the age-period-cohort model and the extended chain-ladder model. *Biometrika* **95**, 987–991.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* **87**, 659–675.

Lee, T. C., Dean, C., and Semenciw, R. (2011). Short-term cancer mortality projections: a comparative study of prediction methods. *Statistics in Medicine* **30**, 3387–3402.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.

Luo, L. (2013). Assessing validity and application scope of the intrinsic estimator approach to the age-period-cohort problem. *Demography* **50**, 1945–1967.

Møller, B., Fekjær, H., Hakulinen, T., Sigvaldason, H., Storm, H. H., Talbäck, M., and Haldorsen, T. (2003). Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches. *Statistics in Medicine* **22**, 2751–2766.

Møller, B., Fekjær, H., Hakulinen, T., Tryggvadóttir, L., Storm, H. H., Talbäck, M., and Haldorsen, T. (2002). Prediction of cancer incidence in the Nordic countries up to the year 2020. *European Journal of Cancer Prevention* **11**, S1–S96.

Møller, B., Weedon-Fekjær, H., and Haldorsen, T. (2005). Empirical evaluation of prediction intervals for cancer incidence. *BMC Medical Research Methodology* **5**, 21.

Plummer, M. (2003). In: Hornik, K., Leisch, F., and Zeileis, A., JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna, Austria.

Qiu, Z., Jiang, Z., Hatcher, J., and Cancer Projection Analytical Network Working Team. (2010a). Comparison of short-term projection methods: proposing the hybrid approach and Bayesian models, Alberta Health Services: 2010-06-30

Qiu, Z., Jiang, Z., Wang, M., Hatcher, J., and the Cancer Projection Analytical Network Working Team. (2010b). Long-term projection methods: comparison of age-period-cohort model-based approaches, Alberta Health Services.

Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* **38**, 556–570.

Riebler, A. and Held, L. (2010). The analysis of heterogeneous time trends in multivariate age-period-cohort models. *Biostatistics* **11**, 57–69.

Riebler, A., Held, L., and Rue, H. (2012a). Estimation and extrapolation of time trends in registry data—borrowing strength from related populations. *Annals of Applied Statistics* **6**, 304–333.

Riebler, A., Held, L., Rue, H., and Bopp, M. (2012b). Gender-specific differences and the impact of family integration on time trends in age-stratified Swiss suicide rates. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **175**, 473–490.

Roos, M., Martins, T. G., Held, L., and Rue, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis* **10**, 321–349.

Rosenberg, P. S. and Anderson, W. F. (2011). Age-period-cohort models in cancer surveillance research: ready for prime time? *Cancer Epidemiology, Biomarkers and Prevention* **20**, 1263–1268.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*, vol. **104**. Chapman & Hall/CRC Press, London.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society—Series B* **71**, 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2016). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*. Conditionally accepted.

Rutherford, M. J., Thompson, J. R., and Lambert, P. C. (2012). Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines. *International Journal of Biostatistics* **8**, 33.

Schmid, V. J. and Held, L. (2007). Bayesian age-period-cohort modeling and prediction—BAMP. *Journal of Statistical Society* **21**, 1–15.

Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics* **22**, 725–734.

Schrödle, B., Held, L., Riebler, A., and Danuser, J. (2011). Using INLA for the evaluation of veterinary surveillance data from Switzerland: a case study. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **60**, 261–279. doi: 10.1111/j.1467-9876.2010.00740.x

Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association* **88**, 355–359.

Seillier-Moiseiwitsch, F., Sweeting, T. J., and Dawid, A. P. (1992). Prequential tests of model fit. *Scandinavian Journal of Statistics* **19**, 45–60.

Simpson, D. P., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2015). Penalising model component complexity: a principled, practical approach to constructing priors (with discussion). *Statistical Science*. To appear.

Smith, T. R. and Wakefield, J. (2016). A review and comparison of age-period-cohort models for cancer incidence. *Statistical Science*. To appear.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science* **333**, 1393–1400.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* **5**, 421–433.

World Health Organization. (2014). Health statistics and information systems, mortality database. URL http://www.who.int/healthinfo/statistics/mortality_rawdata/en/. Accessed on April 25, 2014.

