Received: 16 February 2022 | Revised: 7 February 2023 | Accepted: 1 March 2023

DOI: 10.1002/bimj.202200046

RESEARCH ARTICLE

Biometrical Journal

Comparison of statistical models to predict age-standardized cancer incidence in Switzerland

Bastien Trächsel | **Valentin Rousson** | **Jean-Luc Bulliard** | **Isabella Locatelli**

Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland

Correspondence
Bastien Trächsel, Center for Primary Care and Public Health (Unisanté), University of Lausanne, Route de Berne 113, 1010, Lausanne, Switzerland.
Email: sui.bastien@gmail.com

Funding information
Krebsliga Schweiz, Grant/Award Number: KFS-4385-02-2018

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

Abstract

This study compares the performance of statistical methods for predicting age-standardized cancer incidence, including Poisson generalized linear models, age-period-cohort (APC) and Bayesian age-period-cohort (BAPC) models, autoregressive integrated moving average (ARIMA) time series, and simple linear models. The methods are evaluated via leave-future-out cross-validation, and performance is assessed using the normalized root mean square error, interval score, and coverage of prediction intervals. Methods were applied to cancer incidence from the three Swiss cancer registries of Geneva, Neuchatel, and Vaud combined, considering the five most frequent cancer sites: breast, colorectal, lung, prostate, and skin melanoma and bringing all other sites together in a final group. Best overall performance was achieved by ARIMA models, followed by linear regression models. Prediction methods based on model selection using the Akaike information criterion resulted in overfitting. The widely used APC and BAPC models were found to be suboptimal for prediction, particularly in the case of a trend reversal in incidence, as it was observed for prostate cancer. In general, we do not recommend predicting cancer incidence for periods far into the future but rather updating predictions regularly.

KEYWORDS

age-period-cohort models, age-standardized cancer incidence, autoregressive integrated moving average, Bayesian age-period-cohort models, generalized linear models, interval score, prediction interval, root mean square error, trend reversal

1 | INTRODUCTION

Cancer has progressively become a major public health problem in Western countries like Switzerland where, each year, about 40,000 new cases are diagnosed and 17,000 induced deaths occur (OFS, 2020). Cancer is the second leading cause of mortality in Switzerland, and the main cause in females aged 25–84 years, and in males aged 45–84 years (OFS, 2020). Prediction of cancer incidence is an important public health issue in modern societies to plan resource allocation and prevention interventions and to help address future cancer research.

Many prediction methods are used in the field of cancer incidence, such as age-period-cohort (APC) (Carstensen, 2007; Holford, 1983; Rutherford et al., 2012), and Bayesian age-period-cohort (BAPC) (Riebler et al., 2012; Riebler & Held, 2017;

Schmid & Held, 2004) models, autoregressive integrated moving average (ARIMA) time series models (Hamilton, 1994), neural networks (Hastie et al., 2009), joinpoint regression (Lerman, 1980), Poisson and negative binomial generalized linear models (GLM) (A. C. Cameron & Trivedi, 2013a), state space models (H. S. Chen et al., 2012; Hamilton, 1994), and the vector autoregressive Hilbert–Huang transform (H. S. Chen et al., 2012). In a recent systematic review of 101 published studies on lung cancer incidence prediction, the most commonly used methods were APC, BAPC, GLM, and joinpoint (Yu et al., 2019). APCs were largely applied to predict the incidence of common cancer sites (Coupland et al., 2010; Møller et al., 2002; Rapiti et al., 2014) as well as BAPC (J. K. Cameron & Baade, 2021; W.-Q. Chen et al., 2011; Shi et al., 2021), GLM (NSW Cancer Institute, 2016; Yang et al., 2005; Zemni et al., 2022), and joinpoint (Lin et al., 2021; Rahib et al., 2021; Wong et al., 2021). Other studies used ARIMA models (Earnest et al., 2019; Li et al., 2022; Tsoi et al., 2017).

Comparisons among prediction models have been carried out in the literature. For example, Møller (2003) compared several variants of APC predicting the incidence of the most frequent cancer sites and concluded that APC models using a power link function were superior to those adopting the canonical logarithmic link. To predict the incidence of lung, bronchus, and trachea cancers, Riebler (2012) compared BAPC with the Lee–Carter model, a model used in the demographic field to predict mortality (Lee & Carter, 1992), showing the superiority of BAPC. ARIMA has been compared to neural networks showing the similar performance (Zheng et al., 2020). Using USA state-level data to predict the incidence of the most frequent cancer sites, Chen (2012) compared the performance of APC, BAPC, joinpoint regression, state space models, and the vector autoregressive Hilbert–Huang transform and concluded the superiority of BAPC, although followed closely by APC and state space models. In another study (Clements, 2005), generalized additive models (GAM, a special case of GLM) have shown a better performance than BAPC to predict the incidence of lung cancer.

These comparisons were in general limited to a few models only (e.g., BAPC vs. Lee-Carter, Riebler et al., 2012; or BAPC vs. GAM, Clements, 2005) or to different variants of the same class of models, for example, APC (Møller et al., 2003), with the notable exception of the U.S. study (H. S. Chen et al., 2012), which, however, omitted important methods such as GLM or ARIMA. In addition, most of these comparisons have been made in situations where the incidence was almost constantly increasing, providing a relatively easy setting for prediction. In recent years in Switzerland, we observed some trend reversals, as the incidence of some cancers (e.g., breast, prostate, and skin melanoma) has begun to stabilize or decrease (OFS, 2020). This phenomenon represents an additional difficulty in predicting incidence trends, and it is of interest to compare the different methods in this context.

In the present paper, we sought to compare a large number of methods for predicting the incidence of the most common cancers, including in situations with a trend reversal, based on cancer incidence data from Switzerland. The methods compared include Poisson GLM, of which the Lee–Carter model and joinpoint regression are special cases, APC and BAPC models, ARIMA, and simple linear regression models. We evaluated prediction performance by repeatedly using leave-future-out cross-validation, that is, dividing the data into a training set (on which the models are fitted) and a test set (on which the predictions are evaluated) and considering all possible period partitions of the data in these two sets. This allowed us to produce many prediction settings (scenarios), including some with a trend reversal at the boundary of the training set.

This paper is organized as follows. The data used for our comparisons are presented in Section 2. Section 3 explains how we evaluated the performance of a prediction method by leave-future-out cross-validation using different criteria. All compared methods are described in Section 4. Our results are detailed in Section 5, and a discussion is proposed in Section 6.

2 | DATA DESCRIPTION

2.1 | Data sources

Our primary data sources were population-based Swiss cancer registries of the cantons of Vaud, Geneva, and Neuchatel. Registries record information on all incident cases of malignant neoplasms occurring in their resident population according to international rules, such as the International Classification of Diseases for Oncology (Fritz et al., 2000). For each incident case, information on the incidence date, the patient's date of birth, and the cancer site are recorded, among others.

We used combined data from the three cancer registries of the cantons of Vaud, Geneva, and Neuchatel because they are the oldest registries in Switzerland. Incidence records were available from 1982 to 2016, the latter year being the most recent with complete data. We considered the five most frequent cancer sites: breast, colorectal, lung, prostate, and skin melanoma and formed a final group including all other sites. Considering separately the two sexes, we thus dealt with

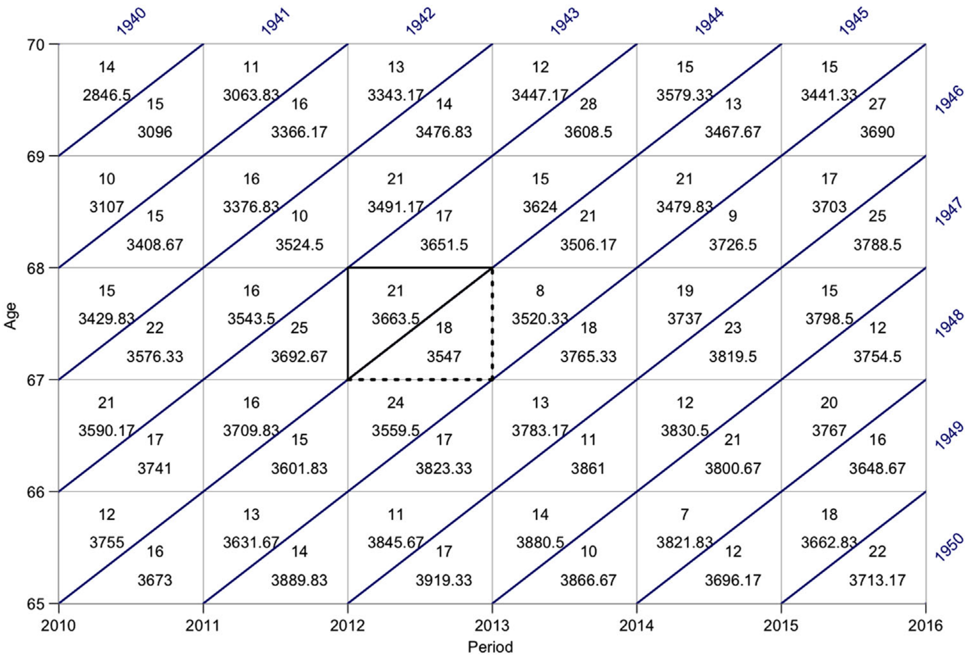

FIGURE 1 Lexis diagram of an extract of the combined data from the Swiss registries of Vaud, Geneva, and Neuchatel. Selected period: January 1, 2010–December 31, 2015. Selected age range: 65th to 70th birthday. In each triangle, the top number represents the number of incident cases $D_{a,p,c}$ and the bottom number corresponds to the person-years $Y_{a,p,c}$ for an age (a), period (p), and birth cohort (c) combination. For example, during 2012 we had $D_{67,2012,1944} = 21$ new cases of breast cancer among women 67 years old who were born in 1944 (triangle in bold) and $D_{67,2012,1945} = 18$ new cases among women of the same age who were born in 1945 (dashed triangle). Corresponding person-years were $Y_{67,2012,1944} = 3663.5$ and $Y_{67,2012,1945} = 3547$, respectively.

five sites for women (breast, colorectal, lung, skin melanoma, and others) and five sites for men (prostate, colorectal, lung, skin melanoma, and others).

Our second source of information was the Swiss Federal Statistical Office (FSO), which produces population figures for each Swiss canton, informing on how many people are alive at the beginning of a calendar year by age and sex.

2.2 | Age, period, and cohort tabulation

Data were aggregated on the three dimensions of age, period, and cohort, using the well-known triangle representation of the Lexis diagram (Figure 1). For each triangle, data include the number of incident cases for the corresponding age a ($a = 0, \dots, 98, 99+$), period p ($p = 1982, \dots, 2016$), and birth cohort c ($c = 1882, \dots, 2016$) combination ($D_{a,p,c}$), and the number of person-years for the same combination ($Y_{a,p,c}$) in the general population, the latter being calculated from population data using the classical method presented in Carstensen (2007). APC-specific incidence rates are then defined by: $\lambda_{a,p,c} = D_{a,p,c}/Y_{a,p,c}$. Each ratio $\lambda_{a,p,c}$ in the Lexis diagram acts as an "observation" in most models considered (see Section 4). Therefore, for each sex and cancer site, we worked with 7000 observations obtained by combining the following factors: 35 periods ($p = 1982, \dots, 2016$), 100 ages ($a = 0, \dots, 98, 99+$), and 2 triangles (upper/lower) separating people in two different birth cohorts (e.g., a person reaching the age of 67 in 2012 may have been born in either 1945 or 1944; see Figure 1). Some of the models below consider only the age and period tabulation:

$$\lambda_{a,p} = \frac{1}{2}\left(\frac{D_{a,p,c}}{Y_{a,p,c}} + \frac{D_{a,p,c+1}}{Y_{a,p,c+1}}\right) = \frac{D_{a,p}}{Y_{a,p}} \text{ (with } c = p - a\text{)}, \tag{1}$$

or only the age and cohort tabulation:

$$\lambda_{a,c} = \frac{1}{2}\left(\frac{D_{a,p,c}}{Y_{a,p,c}} + \frac{D_{a,p+1,c}}{Y_{a,p+1,c}}\right) = \frac{D_{a,c}}{Y_{a,c}} \text{ (with } p = c + a\text{)}. \tag{2}$$

TABLE 1 Summary of quantities related to the triangle tabulation of the Lexis diagram.

a	Age
p	Period
c	Cohort
$D_{a,p}$ ($D_{a,c}$; $D_{a,p,c}$)	Number of new cancer cases for age a and period p (respectively, age a and period c; or age a, period p, and cohort c)
$Y_{a,p}$ ($Y_{a,c}$; $Y_{a,p,c}$)	Person-years for age a and period p (respectively, age a and period c; or age a, period p, and cohort c)
$\boldsymbol{\lambda}_{a,p}$ ($\boldsymbol{\lambda}_{a,c}$; $\boldsymbol{\lambda}_{a,p,c}$)	Incidence rates for age a and period p (respectively, age a and period c; or age a, period p, and cohort c)
Y^*_a	Person-years for age a of a reference population
$\boldsymbol{\lambda}^*_p$	Standardized incidence rates of year p

2.3 | Standardized incidence rates

To compare incidence rates over periods using a single quantity, age-specific incidence rates should be combined taking into account changes in the population structure over time. This can be achieved by weighting the age-specific incidence rates of a period p ($\lambda_{a,p}$) for the age distribution of a reference population Y^*_a ($a = 0, \ldots, 99+$), resulting in so-called *standardized incident rates*:

$$\lambda^*_p = \frac{\sum_{a=0}^{99+} \lambda_{a,p} Y^*_{\mathrm{a}}}{\sum_{a=0}^{99+} Y^*_{\mathrm{a}}}. \tag{3}$$

It has been shown (Spiegelman & Marks, 1966) that when comparing incidence rates over time, the choice of the reference population has little impact on the results. The same is true when making predictions. In this study, we chose the population of the year 2000 of the aggregated cantons of Vaud, Geneva, and Neuchatel as the reference for standardization. A summary of the notations used for quantities related to the Lexis diagram is provided in Table 1.

The standardized incidence rates λ^*_p ($p = 1982, \ldots, 2016$) can be seen in Figure 2 for each sex and cancer site. While some rates show a monotonic trend, for example, lung cancer, which continues to increase for women and decrease for men at the present time, other rates show a recent stabilization (breast cancer, skin melanoma) or a trend reversal, in particular prostate cancer, whose incidence began to decrease in the 2000s after decades of increase.

While the majority of the models considered in this paper use tabulated data by age, period, and/or cohort, predicting specific incidence rates $\lambda_{a,p}$, $\lambda_{a,c}$ or $\lambda_{a,p,c}$, we followed the literature (H. S. Chen et al., 2012; Clements, 2005; Møller et al., 2003) and compared their prediction performance using the standardized incidence rates λ^*_p (3). This is the most widely used choice when predicting the cancer burden (Møller et al., 2002; Rapiti et al., 2014) and allows the comparison of the predictions by ARIMA and simple linear regression models which model directly standardized incidence rates (see Section 4).

3 | EVALUATING PREDICTION PERFORMANCE

3.1 | Leave-future-out cross-validation

Models were compared by repeatedly applying the principle of leave-future-out cross-validation (Bürkner et al., 2020). The procedure can be summarized as follows. *First*, for a given sex/cancer site combination, chose a cutoff time t in the range: $t = 2001, \ldots,\ 2015$. Given the cutoff, the leave-future-out cross-validation consists in (a) fitting a model on incidence rates from $T_0 = 1982$ until year t (*training set*), (b) predicting incidence rates for the second part of the data (*test set*), that is, from the year $(t + 1)$ until the last year available $T = 2016$, and (c) evaluating the prediction performance from $(t + 1)$ to T according to criteria below. *Second*, repeat the procedure moving t in the range $t = 2001, \ldots,\ 2015$ and for each sex/cancer site combination. The range for the cutoff t is chosen to ensure having at least 20 observations in the training set and at least one observation in the test set.

Because we had five cancer sites for each sex and 15 possible cutoffs for each sex/cancer site combination ($t = 2001, \ldots,\ 2015$), we obtained 150 different scenarios (see Figure 3). All models presented in Section 4 will be fitted on each of these 150 scenarios.

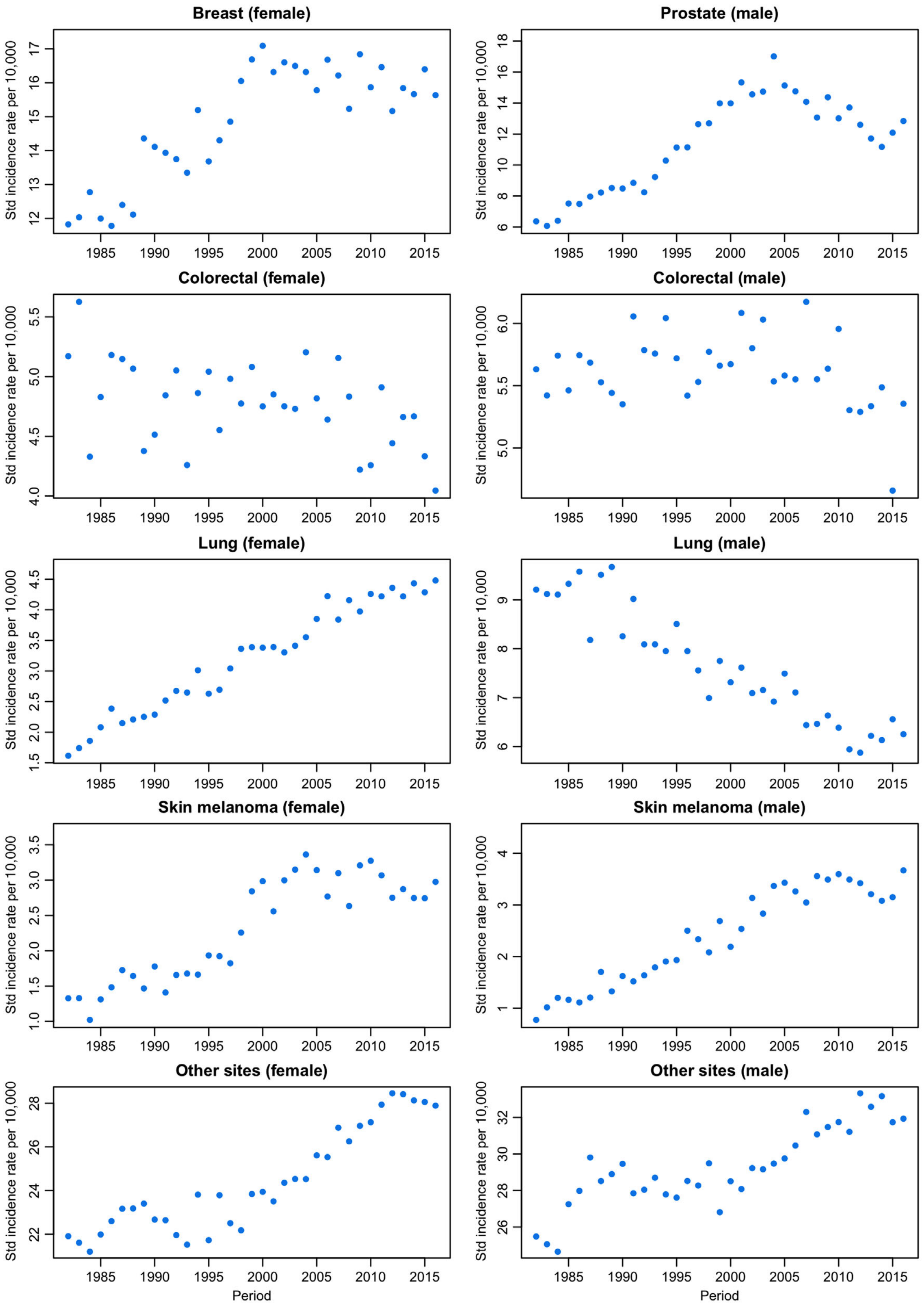

FIGURE 2 Standardized incidence rates per 10,000 inhabitants over the period 1982–2016 for the most common cancer sites and the two sexes. Combined data from the Swiss registries of Vaud, Geneva, and Neuchatel. Reference for standardization: combined population of Vaud, Geneva, and Neuchatel in 2000.

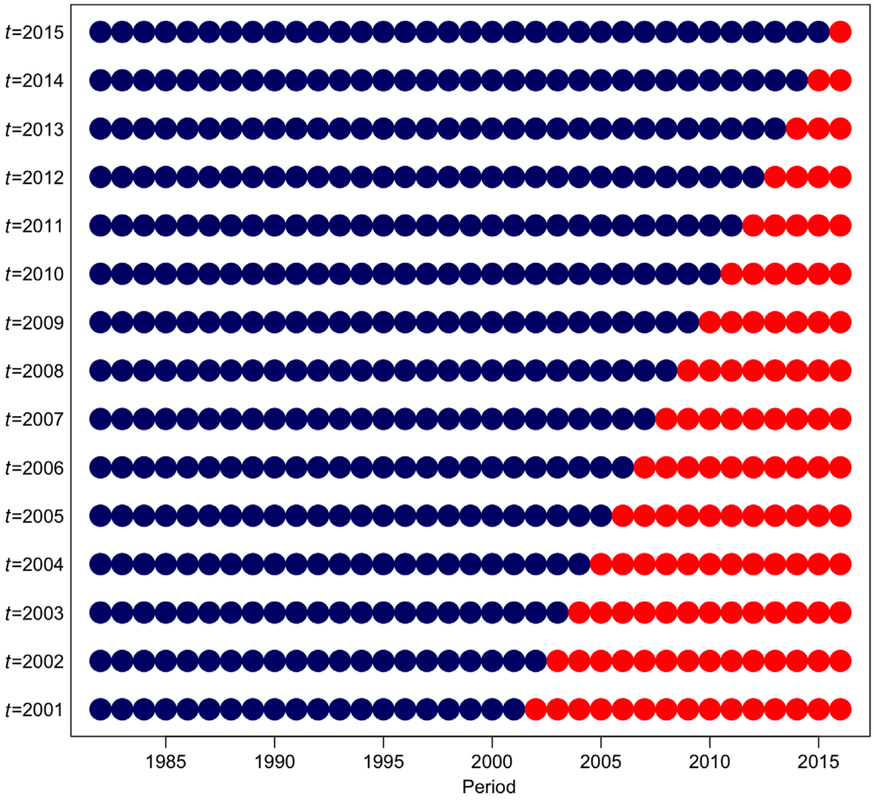

FIGURE 3 Illustration of the 15 leave-future out cross-validation scenarios for each sex and cancer site, based on partitioning the time range into a training set (blue dots) and a test set (red dots) according to a cutoff time t (working with five cancer sites for women and five for men resulted in 150 different scenarios).

3.2 | Comparison criteria

Our main criterion of prediction performance was the normalized root mean squared error (NRMSE) (Hyndman & Koehler, 2006), a measure of point prediction accuracy. This criterion takes the following form in each scenario, that is, for a given cutoff time t, $t = 2001, \ldots, 2015$, and for each sex/cancer site combination:

$$\mathrm{NRMSE}\,(t) = \frac{\sqrt{\sum_{x=t+1}^{T} \left(\hat{\lambda}_x^* - \lambda_x^*\right)^2 / \left(T-t\right)}}{\sum_{x=t+1}^{T} \lambda_x^* / \left(T-t\right)}. \tag{4}$$

Here λ_x^* is the observed (i.e., actual) standardized incidence rates for the period x in the test set and for the considered sex/cancer site combination, and $\hat{\lambda}_x^*$ the corresponding predicted standardized incidence rate. The denominator of (4) represents the mean standardized incidence rate in the test set, allowing normalizing of the statistics. The prediction performance (prediction accuracy) of each model was evaluated by the mean NRMSE (M-NRMSE) across the 150 scenarios. A smaller M-NRMSE value indicates better accuracy, with a minimum of 0. We also looked at three time horizons for prediction, assessing separately the short-term M-NRMSE (1–5 years, $t+1 \leq\ x \leq t+5$), medium-term M-NRMSE (6–10 years, $t+6 \leq x \leq t+10$), and long-term M-NRMSE (11–15 years, $x \geq t+11$). This represented 150, 100, and 50 scenarios, respectively.

In a first sensitivity analysis, we considered an alternative of (4), the normalized mean absolute error (NMAE), defined as follows for a given cutoff time, $t = 2001, \ldots, 2015$, and for each sex/cancer site combination (Hyndman & Koehler, 2006):

$$\mathrm{NMAE}\,(t) = \frac{\sum_{x=t+1}^{T} |\widehat{\lambda}_x^{\,*} - \lambda_x^*| / (T-t)}{\sum_{x=t+1}^{T} \lambda_x^* / \left(T-t\right)}, \tag{5}$$

and considered the mean NMAE (M-NMAE) across the 150 scenarios. In a second sensitivity analysis, we considered the median (instead of the mean) of the NRMSE and the NMAE across scenarios (Med-NRMSE and Med-NMAE).

As an additional comparison criterion, we evaluated the quality of the 95% prediction intervals provided by the different methods to inform on the prediction uncertainty. For this, we used the coverage rate (CR) and the interval score (IS) (Gneiting et al., 2007). The CR measures the empirical coverage of a prediction interval, that is, the probability that it contains the actual standardized incidence rate and is defined as follows for a given cutoff time $t = 2001, \dots, 2015$, and for each sex/cancer site combination:

$$CR\left(t\right) = \frac{\sum_{x=t+1}^{T} I(L_x < \lambda_x^* < U_x)}{(T-t)}, \tag{6}$$

where L_x and U_x are the lower and upper bounds of a 95% prediction interval calculated for period x $(x = t+1, \dots, T)$ and I is an indicator function. The IS is a strictly proper scoring rule, which is related to the width of a prediction interval, with a penalty in cases where the interval does not contain the actual standardized incidence rate, and is defined as follows for a given cutoff time $t = 2001, \dots, 2015$, and for each sex/cancer site combination (Gneiting et al., 2007):

$$IS\left(t\right) = \frac{\sum_{x=t+1}^{T} \left\{ (U_x - L_x) + \frac{2}{0.05} \left[(L_x - \lambda_x^*) I\left(\lambda_x^* < L_x\right) + (\lambda_x^* - U_x) I\left(\lambda_x^* > U_x\right) \right] \right\}}{(T-t)}. \tag{7}$$

The CR of a prediction interval should be as close to 95% as possible, while the IS should be as small as possible with a minimum of 0. As for our primary criteria, CR and IS were averaged across the 150 scenarios, obtaining a mean CR and a mean IS (M-CR and M-IS).

Finally, we reported the percentage of scenarios for which the fitting algorithms used by the different methods converged (i.e., they were able to produce predictions), which was not always 100%. Note that all the above criteria were calculated excluding the few cases where the computation did not converge.

All models below will be evaluated and compared according to all criteria (M-NRMSE, med-NMRSE, M-NMAE, Med-NMAE, M-CR, M-IS, and convergence rate). All detailed results are given in the Appendix of the Supporting Information.

4 | PREDICTION METHODS

4.1 | Generalized linear models

We considered the Poisson GLM as presented in Chapter 2.4 of A. C. Cameron and Trivedi (2013a). The number of cases $D_{a,p}$ $(D_{a,c})$ in a certain area of the Lexis diagram (see Figure 1) is assumed to follow a Poisson distribution whose mean $\lambda_{a,p} \cdot \mathrm{Y}_{a,p}$ $(\lambda_{a,c} \cdot \mathrm{Y}_{a,c})$ depends (usually via a log link) on the age and period (or age and cohort). The person-years $Y_{a,p}$ $(Y_{a,c})$ is an offset of the model. Due to the *age = period−cohort* relationship, the three variables cannot be introduced simultaneously in a model, so that we considered data either in squares of the Lexis diagram (age and period) or in parallelograms (age and cohort, see Figure 1), that is, only two effects were considered at a time, with a possible interaction between the two:

$$\log\left(\frac{D_{a,z}}{Y_{a,z}}\right) = \log\left(\lambda_{a,z}\right) = ns_k\left(a\right) + f\left(z\right) + g\left(a\right) \cdot h\left(z\right) \tag{8}$$

with $z = \{p, c\}$; $f, g, h = \{\emptyset, id, ns_{\mathrm{k}}\}$; $k = 1, \dots, 4$. The age a was introduced into the model with natural splines ns_k, that is, cubic splines based on k knots, linear outside the extreme knots (Ruppert et al., 2003). The period (or the cohort) was either absent $(f = \emptyset)$, either introduced linearly, that is, via an identity function $(f = id)$ or by splines $(f = ns_k)$. Both age and period (or age and cohort) were introduced either linearly or by splines in the interaction term, with the constraint that the degrees of freedom for the period (or the cohort) effect must be lower or equal in the interaction term than in the main effect. So, if the period (or the cohort) was introduced linearly as the main effect $(f = id)$, it was not introduced via splines in the interaction $(h \neq ns_k)$, and if the period (or the cohort) was not introduced as a main effect $(f = \emptyset)$, it was not in an interaction term $(h = \emptyset)$. For a given number of knots k, this defined five models without interaction and 12 models with interaction. Varying the number of knots between $k = 1$ and $k = 4$ (beyond which we encountered problems of lack of convergence of the fitting algorithm), we obtained $4 \times (5 + 12) = 68$ GLM. The Lee–Carter model (1992), a wellestablished model used in demography to predict mortality rates, is one of those 68 models where the age and the period are introduced with splines, with an interaction between the two.

Finally, we considered a GLM obtained via a model selection strategy that consists in selecting among the 68 GLM models the one that best fits the data in the training set according to the Akaike information criterion (AIC), so that a possibly different GLM is selected in each scenario. This corresponds to a 69th GLM that we call GLM AIC.

The well-known joinpoint regression (Lerman, 1980) was included as a 70th GLM. This is an age-period model with the period introduced with linear splines, with knots representing times where a change of trend occurs. Unlike classical splines, where the knots are chosen a priori, knots in joinpoint regression are determined from the data according to AIC. We only considered a single knot joinpoint model.

Predictions of incidence rates were obtained by extrapolating period, respectively, cohort effects. For age-cohort models, cohort effects must be extrapolated for future periods, with the particularity of requiring the prediction of the effect of some new cohorts and the removal of the effect of old cohorts, as each year some cohorts exit and new cohorts enter. Predictions of $\lambda_{a,c}$ or $\lambda_{a,p}$ ($\hat{\lambda}_{a,c}$ or $\hat{\lambda}_{a,p}$) were then aggregated and standardized according to (3) to obtain predicted standardized incidence rates $\hat{\lambda}^*_p$. The variance of $\hat{\lambda}_{a,c}$ (or $\hat{\lambda}_{a,p}$) was estimated by the delta method as in Chapter 3 of A. C. Cameron and Trivedi (2013b), and prediction intervals were obtained based on this variance following the implementation provided in the `R package trending` (Schumacher & Jombart, 2021).

4.2 | Age-period-cohort models

APC models, introduced by Holford (1983), are Poisson GLM allowing to include simultaneously age, period, and cohort effects. In order to make the model identifiable despite the linear relationship between the three variables, some constraints are added (Carstensen, 2007). Whereas age, period, and cohort were originally considered as factors, forcing a coarse tabulation of 5-year groups (Clayton & Schifflers, 1987; Holford, 1983; Møller et al., 2002), more recent developments in APC, which we followed in our study, adopted splines to model three effects (Carstensen, 2007; Carstensen et al., 2022). With the canonical log link of Poisson GLM, a spline APC model takes the following form:

$$\log\left(\lambda_{a,p,c}\right) = ns_k\left(a\right) + ns_k\left(p\right) + ns_k\left(c\right) + \text{drift}. \tag{9}$$

Here the drift is a linear trend effect, ascribed to both period and cohort and $ns_k(p)$ (respectively, $ns_k(c)$) are residual nonlinear effects specific to the period (respectively, the cohort). For these splines, we considered between k =3, 4, or 5 knots. In addition to the canonical log link, other link functions (Dunn & Smyth, 2018) have been compared in the literature for their prediction performance, where the 1/5 power link performed best (Møller et al., 2003). For this reason, we have adopted the 1/5 power link in our comparisons, as an alternative to the log link (9).

Two options have been adopted in the literature to extrapolate the incidence rates $\hat{\lambda}_{a,p,c}$ for future periods and cohorts. The most common one consists in extrapolating only the drift (Yu et al., 2019); the alternative consists in adding a linear extrapolation of the nonlinear effects (Rutherford et al., 2012). Both options are compared in our study. In what follows, the two extrapolation strategies are referred to as *drift only*, and *all effects*, respectively. Varying the number of knots between k =3, 4 or 5, and considering two link functions and two extrapolation strategies, we obtained $3 \times 2 \times 2 = 12$ APC models.

As with GLM, the predictions $\hat{\lambda}_{a,p,c}$ were finally aggregated and standardized as in (3) to produce predicted standardized incidence rates $\hat{\lambda}^*_p$ and prediction intervals.

4.3 | Bayesian age-period-cohort models

A BAPC model (Riebler et al., 2012; Riebler & Held, 2017; Schmid & Held, 2004) has the same formulation as an APC model, the difference being how age, period, and cohort effects are included in the model and how they are fitted. Instead of using splines, Bayesian models are based on the first- or second-order random walk (RW_1 or RW_2) specification for each effect of the model:

$$\log\left(\lambda_{a,p,c}\right) = RW_l\left(\text{a}\right) + RW_l\left(p\right) + RW_l\left(c\right), \tag{10}$$

where $l = 1, 2$ is the order of the random walk. Considering, for instance, the period effect (the same holds for age and cohort effects), a Bayesian first-order random walk takes the following form:

$$RW_1\,(p) = \Delta p \sim N\left(0, \sigma_1^2\right), \tag{11}$$

where Δp is the difference of the last two period effects. With the RW_1 model, each (age, period, and cohort) effect is taken as being equal to the previous one plus a random number from a normal distribution centered on zero, with variance σ_1^2. The latter acts as a smoothing parameter: with an infinite variance, the fitted model will pass through all the data points in the training set, while a small prior variance will give rise to a smoother fit. We considered different choices for the hyper-prior distribution of parameter σ_1^2. Our first option was the most commonly adopted distribution, that is, a flat (noninformative) gamma(1, 5e−5) for the three effects of age, period, and cohort (Riebler & Held, 2017). As alternative options, we considered a tighter gamma(1, 5e−3), a larger gamma(1,5e−7), and PC(u=1,α=0.01) (Lindgren & Rue, 2015) priors for the three effects, as well as an option with a gamma(1, 9e−4) for the age effect and gamma(1, 2.5e−4) for the period and cohort effects, as in Riebler and Held (2017). Considering that we have 7000 observations in the Lexis diagram, the hyper-prior has a limited effect. When making predictions from a Bayesian RW_1 model, the last fitted period and cohort effects are extrapolated as constant.

Considering again the period effect as an example, a Bayesian second-order random walk takes the form:

$$RW_2\,(p) = \Delta^2 p \sim N\left(0, \sigma_2^2\right), \tag{12}$$

where $\Delta^2 p$ is the second difference (difference of the difference) involving the last three period effects. In a RW_2 model, each age, period, and cohort effect is based on the two previous effects and adding a random number from a normal distribution centered on zero, with variance σ_2^2. The latter acts again as a smoothing parameter: the smaller the prior variance, the smoother the fit. The same five priors adopted for σ_1^2 were adopted for σ_2^2. In a Bayesian RW_2 model, the future period and cohort effects are predicted by a linear trend based on the last two fitted effects. Compared to an RW_1 model, an RW_2 model will be generally smoother.

Since we considered five priors for each of the RW_1 and RW_2 models, we included a total of 10 BAPC in our comparison. As with GLM and APC models, the predicted rates $\hat{\lambda}_{a,p,c}$ from a BAPC model should be combined to obtain predictions of standardized incidence rates $\hat{\lambda}^*_p$. Bayesian RW_1 and RW_2 models were fitted using the `INLA` (Lindgren & Rue, 2015) and `BAPC packages` (Riebler & Held, 2017) from `R` software (R Core Team, 2022) to get point predictions and prediction intervals, respectively. The `package INLA` is based on the works of Rue (2009), Martins (2013), and Lindgren (2011, 2008, 2015).

4.4 | ARIMA time series models

ARIMA time series models are econometric models defined by combining a difference autoregressive model with a moving average model (Hamilton, 1994). Let $\Delta^d \lambda^*_p$ be the standardized rates λ^*_p differenced d times. The ARIMA(h,d,q) is expressed as

$$\Delta^d \lambda^*_p = \alpha_0 + \alpha_1 \Delta^d \lambda^*_{p-1} + \alpha_2 \Delta^d \lambda^*_{p-2} + \cdots + \alpha_h \Delta^d \lambda^*_{p-h} + \epsilon_p + \theta_1 \epsilon_{p-1} + \theta_2 \epsilon_{p-2} + \cdots + \theta_q \epsilon_{p-q}, \tag{13}$$

where ϵ_p are normally distributed residuals, α_1,…,α_h are the coefficients of the autoregressive (AR) part of the model, $\theta_1, \ldots, \theta_q$ are the coefficients of the moving average (MA) part, and α_0 is a constant. In an ARIMA(h,d,q) the predictors are lagged h data points for the autoregressive part and q residuals are considered for the moving average part, which are all d differenced. We considered all orders from (0,0,0) to (3,3,3) for (h,d,q) which represents $4^3 = 64$ different models. As for GLM, we finally considered an ARIMA obtained via a model selection strategy that consists in selecting among the 64 ARIMA models the one that best fits the data in the training set according to the AIC, so that a possibly different ARIMA is selected in each scenario. This corresponds to the 65th ARIMA method that we call ARIMA AIC.

Unlike the models presented in the previous sections, the predictions obtained with an ARIMA model are based directly on an extrapolation of the standardized rates (without the need for any aggregation). We used the function ARIMA from R (R Core Team, 2022) to fit the models, predict the incidence rates, and compute the prediction intervals.

4.5 | Linear models

In order to get another standard comparator for the above methods, we finally considered a simple linear model (LM) fitted on the last r data points of the training set:

$$\lambda_p^* = \alpha + \beta p + \epsilon_p, \tag{14}$$

where $p = (t - r + 1), \ldots, t$, ϵ_p are normally distributed residuals and α and β are the intercept and slope of the model, respectively. Letting r vary between 3 and 10 data points, we considered eight LM. In these models, predictions of standardized incidence rates are made by extrapolating the fitted trend, and prediction intervals are obtained in the classical framework of regression models.

In summary, considering the five classes of models described in this section, we compared 165 different models: 70 GLM, 12 APC, 10 BAPC, 65 ARIMA, and 8 LM. An exhaustive list of all the models compared can be found in Figures A1–A4 of the Supporting Information Appendix.

5 | RESULTS

All models described in Section 4 were fitted to the 150 available scenarios detailed in Section 3.1 and evaluated according to the performance criteria presented in Section 3.2. For each model, the NRMSE and IS distribution over the 150 scenarios is given in Figures A1–A8 of the Supporting Information Appendix. Most of these distributions show large variability and are highly skewed, with M-NRMSE > Med-NRMSE, as M-NRMSE strongly penalizes the poor performance of some models in particularly difficult scenarios, for example, in the case of a trend reversal (and the same is true for the M-IS).

5.1 | Illustration of two selected scenarios

To illustrate the heterogeneity of predictions among models, we have plotted in Figure 4 two selected scenarios, one for female lung cancer and one for (male) prostate cancer, both with a cutoff time of $t = 2010$. In the former, incidence rates are always increasing, representing a simple setting for predictions, while in the latter the incidences begin to decrease around 2005, giving a more challenging situation for predictions. It can be seen that in the former scenario almost all models predict similar incidence rates, whereas in the latter the heterogeneity among predictions is impressive, as a trend reversal occurs towards the end of the training set. In this setting, some models adapt to the new trend (e.g., an APC extrapolating *all effects*, represented in green), while others fail to adapt (e.g., an APC extrapolating the *drift only*, represented in dark cyan). We can observe that highly flexible models, although fitting closely to the data points in the training set, are not necessarily the best for prediction in the test set (e.g., ARIMA (3,3,0) in red).

The Supporting Information Appendix contains more detailed results for prostate cancer. Figure A9 shows the predictions of the different models for each of the 15 scenarios (i.e., varying the cutoff time t) for this cancer site. When the cutoff time occurs before the trend reversal, all models predict excessively high incidence rates, but some models adapt quicker than others when the cutoff time occurs during or just after the trend reversal. For each model, Figures A10–A13 give the NMRSE distribution over the 15 scenarios for prostate cancer, showing a generally worse performance of the methods (higher NMRSE values and more skewed distribution) than what we had in Figures A1–A4 (when considering all cancer sites). The NRMSE and IS distributions of each model over the 15 scenarios for all 10 sites considered in this study can be found in the Supporting Information. In what follows, we discuss and compare the prediction performance achieved by the different models.

5.2 | Comparison within classes of models

In the class of GLM (Section 4.1), recall that we varied (a) the number of knots k of the splines between 1 and 4, and (b) the variables included in the model and the functional forms used. Models with $k = 3$ and $k = 4$ knots achieved smaller M-NMRSE (Figure 5a). Models including age and period performed better than models including only age or age and cohort, with better performance when the period was included via splines rather than just linear, while models including

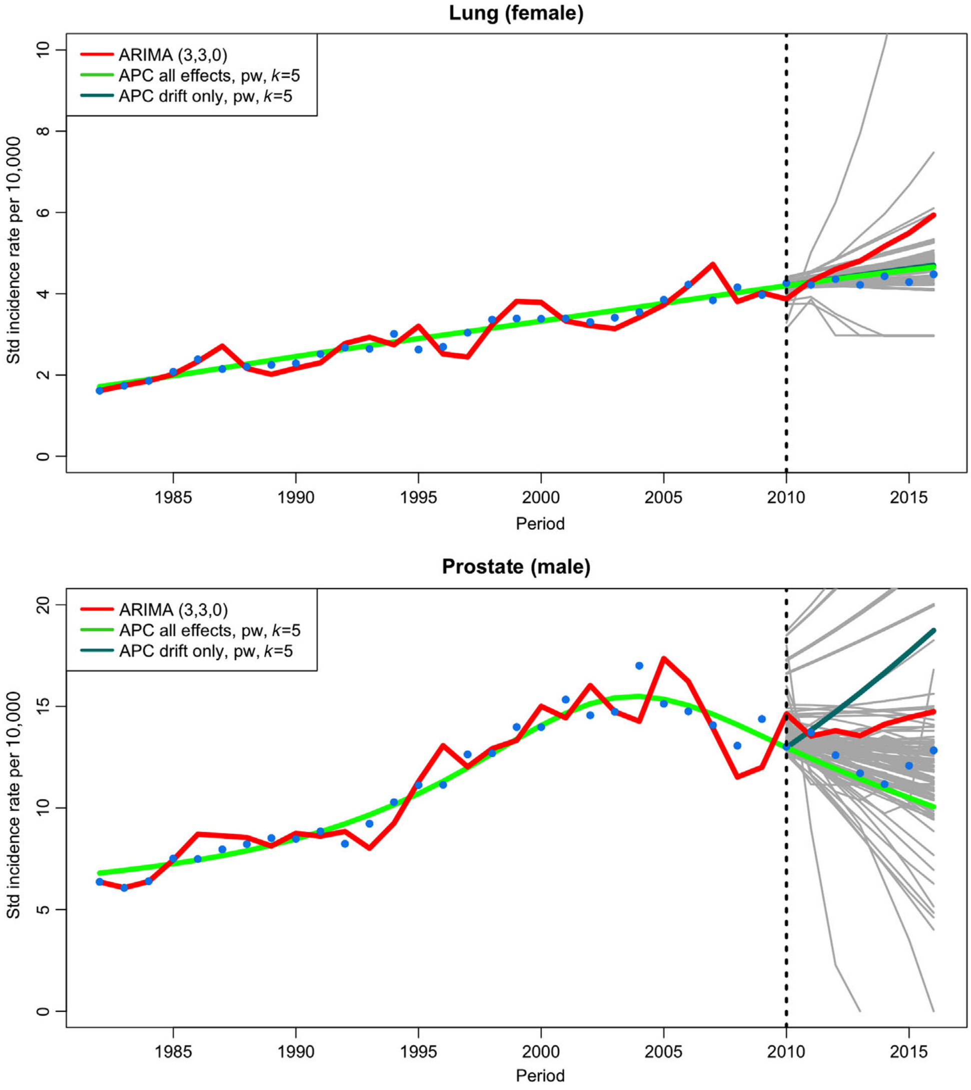

FIGURE 4 Illustration of predictions made by the different models in one of the 15 leave-future-out cross-validation scenarios (training set 1982–2010; test set 2011–2016) for lung (female) and prostate (male) cancer standardized incidence rates. Combined data from the Swiss registries of Vaud, Geneva, and Neuchatel, 1982–2016. Three models are highlighted (APC extrapolating only the drift, APC extrapolating the drift + all nonlinear effects, and ARIMA (3,3,0)). All other models are in gray. APC, age-period-cohort; ARIMA, autoregressive integrated moving average.

an interaction term performed similarly to those without interaction (Figure 5b). The best GLM with $k = 4$ knots was a model including age and period, along with a linear interaction between the two variables (M-NRMSE = 0.138), while the best GLM with $k = 3$ knots was a model including age and period without interaction (M-NRMSE = 0.143), the latter being simpler and converging more often than the former (100% vs. 96%). Both models performed much better than the more complex Lee–Carter model (M-NRMSE = 0.180). On the other hand, the use of AIC for selecting the GLM showed an extremely poor performance, as did the joinpoint model (both M-NRMSE > 7; see also Figures A1 and A2 in the Supporting Information Appendix).

In the APC class of models (Section 4.2), we varied the number of knots k between 3 and 5, while considering two possible link functions (log or 1/5 power), and two different extrapolation strategies (*drift only* or *all effects*). Results are summarized in Figure 5c. We found that using the 1/5 power link improved the predictions compared to using the canonical log link, especially when extrapolating the drift only. The optimal number of knots was $k = 3$ when using the first extrapolation strategy (*drift only*) and $k = 5$ when using the second (*all effects*). In general, the second extrapolation strategy improved predictions compared to the first. The best APC model was thus a model using the 1/5 power link, with five knots and extrapolating *all effects* (M-NRMSE = 0.140; see also Figure A3 in the Supporting Information Appendix).

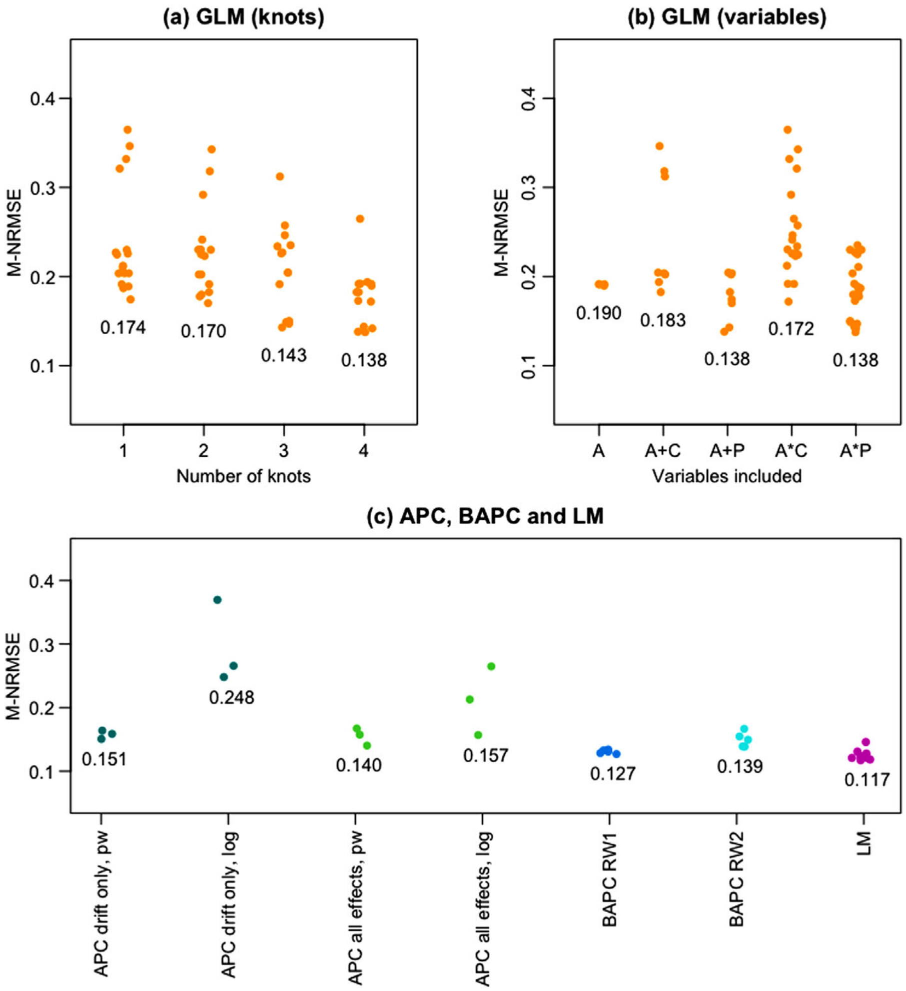

FIGURE 5 Mean normalized root mean square error (M-NRMSE) for GLM (a) and (b); APC, BAPC, and LM (c) across 150 scenarios (15 leave-future-out cross-validation scenarios for each sex and five cancer locations per sex) obtained on combined data from the Swiss registries of Vaud, Geneva, and Neuchatel, 1982–2016. In the labels, "A" indicates GLM only including age; "A+P" (respectively, "A+C") indicates GLM including age and period (respectively, age and cohort) without interaction; "A · P" ("A · C") indicates GLM including age, period (respectively, age, cohort) with interaction. For APC, "drift only" and "all effects" refer to prediction strategies extrapolating only the drift, respectively, the drift + all nonlinear effects, "log" indicates the logarithmic link, and "pw" is the 1/5 power link. For BAPC, RW_1 and RW_2 refer to first, and second-order random walk. APC, age-period-cohort; BAPC, Bayesian age-period-cohort; GLM, generalized linear models; LM, linear model.

In the BAPC class of models (Section 4.3), the choice of the priors for the period, cohort, and age effects had almost no impact on the M-NRMSE, while (using a PC hyper-prior) the performance was better for the RW_1 model (M-NRMSE = 0.127) than for the more complex RW_2 model (M-NRMSE = 0.139), as summarized in Figure 5c (see also Figure A3 in the Supporting Information Appendix).

In the ARIMA class of models (Section 4.4), we varied the autoregressive, integrated, and moving average orders from 0 to 3. Results are summarized in Figure 6. Changing the autoregressive and the moving average order did not have much impact on the prediction performance. On the other hand, working with difference data did improve the performance up to order 1, but got worse beyond. The smallest M-NRMSE was achieved by an ARIMA (2,1,1) (M-NRMSE = 0.078), with a convergence of 96%, while the best ARIMA with a convergence of 100% was ARIMA (1,1,0) (M-NRMSE = 0.084). As for GLM, using the AIC criterion to select a possibly different ARIMA model depending on the scenario resulted in a worse prediction performance (M-NRMSE = 0.119) than when systematically opting for an ARIMA (2,1,1) or (1,1,0) (see also Figure A4 in the Supporting Information Appendix).

Finally, in the class of LM (Section 4.5), the best performance was achieved using the last $r = 7$ data points of the training set to fit the model (M-NRMSE = 0.117; see Figure 5c and Figure A3 of the Supporting Information Appendix).

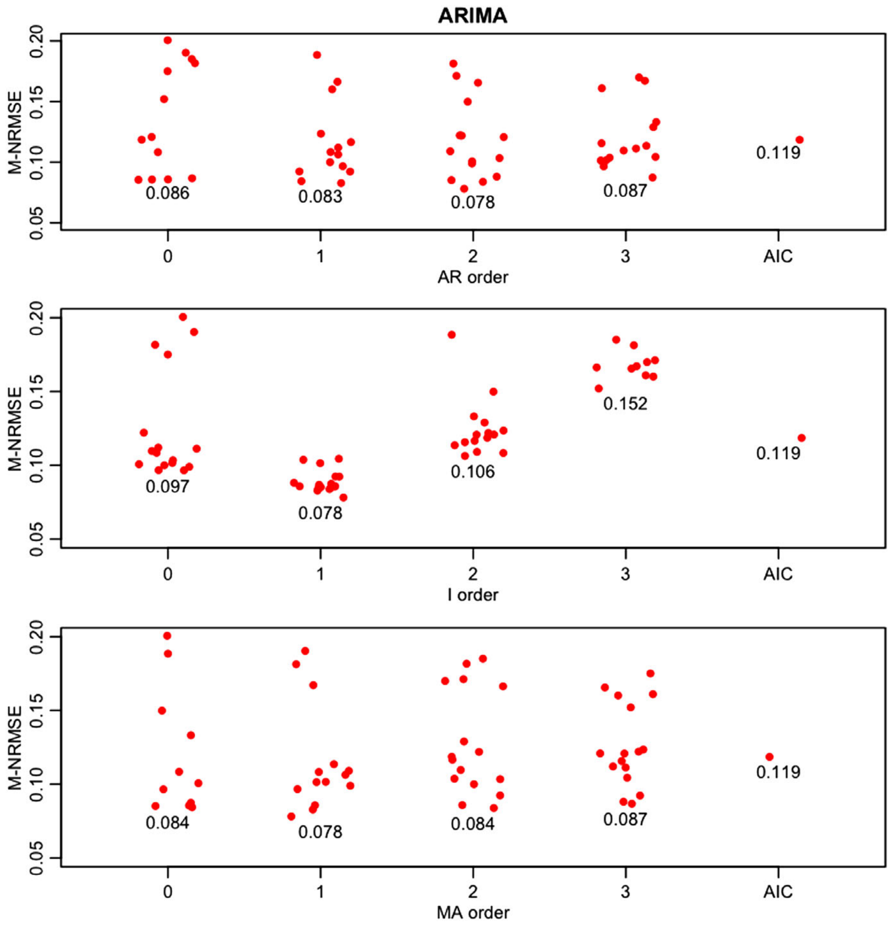

FIGURE 6 Mean normalized root mean square error (M-NRMSE) for ARIMA with orders of each component (AR, I, and MA) between 0 and 3, and an ARIMA AIC across 150 scenarios (15 leave-future-out cross-validation scenarios for each sex and five cancer locations per sex) obtained on combined data from the Swiss registries of Vaud, Geneva, and Neuchatel, 1982–2016. AIC, Akaike Information Criterion; AR, autoregressive; ARIMA, autoregressive integrated moving average; MA, moving average.

The Supporting Information Appendix also contains the same summary results as in Figures 5 and 6 for the M-IS criterion (Figures A14 and A15). Within-class comparisons in terms of M-IS are largely consistent with that obtained using M-NMRSE. The best GLM was a model with three knots including age and period with an interaction (M-IS = 0.199; Figures A14, A5, and A6). The best APC was a model using the 1/5 power link, with five knots and extrapolating *all effects* (M-IS = 0.504; Figures A14 and A7). The best BAPC was a RW_1 model using a gamma(1, 2.5e−4) prior (M-IS = 0.200; Figures A14 and A7). In the ARIMA class, the best performance was achieved by ARIMA(1,1,0) (M-IS = 0.098; Figures A15 and A8), while in the class of LM the best performance was obtained using the last $r = 4$ data points (M-IS = 0.181; Figures A14 and A7).

5.3 | Comparison between classes of models

The performance of the best models from each class identified above based either on M-NRSME or on M-IS, as well as of some other well-known models, is summarized in Table 2, where models are ordered in terms of M-NRSME. Based on this criterion, simple ARIMA models, such as (2,1,1) and (1,1,0) and LM using the last seven data points, performed better than the best models among the more complex (GLM, APC, and BAPC) classes of models. Looking at detailed results in Supporting Information Appendix (Figures A1–A4), one can see that ARIMA models with orders up to (3,1,3) as well as LM models using the last 5–10 data points were also ahead of the other methods. The AIC-based ARIMA outperformed GLM, APC, and BAPC models, while being inferior to simple ARIMA and LM. Finally, the single knot joinpoint and Lee–Carter models showed extremely poor performance.

TABLE 2 Prediction performance for the best models within each class (GLM, APC, BAPC, ARIMA, and LM) according to the mean normalized root mean square error (M-NRMSE) and to the mean interval score (M-IS), and for some other selected models, across 150 scenarios (15 leave-future-out cross-validation scenarios for each sex and five cancer locations per sex) obtained on combined data from the Swiss registries of Vaud, Geneva, and Neuchatel, 1982–2016. Also given is the performance of short-term, medium-term, and long-term predictions provided by M-NRMSE (1–5 years), M-NRMSE (6–10 years), and M-NRMSE (11–15 years). Alternative criteria include the median NMRSE (Med-NRMSE), the mean and median normalized mean absolute error (M-NMAE and Med-NMAE), and the mean coverage rate (M-CR). The last column indicates the percentage of convergence of the methods over the 150 scenarios. Models in this table are ordered according to M-NRMSE.

	M-NRMSE	1–5 years	6–10 years	11–15 years	Med-NRMSE	M-NMAE	Med-NMAE	M-IS	M-CR	Convergence
ARIMA (2,1,1)	0.078	0.068	0.090	0.120	0.078	0.069	0.065	0.103	93.3	96
ARIMA (1,1,0)	0.084	0.071	0.101	0.141	0.079	0.075	0.069	0.098	96.7	100
LM $r = 7$	0.117	0.079	0.153	0.280	0.076	0.103	0.067	0.234	88.8	100
ARIMA AIC	0.119	0.087	0.150	0.223	0.089	0.105	0.079	0.258	78.3	100
BAPC RW1 PC(u=1,α=0.01) prior	0.127	0.085	0.175	0.297	0.083	0.111	0.076	0.205	98.3	89
BAPC RW1 gamma(1,5e-3)	0.129	0.086	0.179	0.303	0.086	0.113	0.074	0.200	99.5	88
LM $r = 4$	0.131	0.091	0.168	0.293	0.083	0.116	0.075	0.181	94.6	100
GLM $ns_4(a) + ns_4(p) + a \cdot p$	0.138	0.087	0.188	0.357	0.088	0.120	0.074	0.377	61.6	96
BAPC RW2 PC prior	0.139	0.084	0.178	0.384	0.091	0.119	0.074	0.337	99.8	86
APC all effects, pw, $k = 5$	0.140	0.088	0.178	0.360	0.093	0.123	0.078	0.504	61.1	97
GLM $ns_3(a) + ns_3(p)$	0.143	0.088	0.186	0.373	0.088	0.124	0.077	0.496	59.8	100
GLM $ns_3(a) + ns_3(p) + ns_3(a) \cdot p$	0.149	0.088	0.194	0.413	0.086	0.128	0.075	0.199	85.5	100
APC drift only, pw $k = 3$	0.151	0.105	0.199	0.303	0.105	0.132	0.094	0.607	51.2	85
GLM Lee Carter, $k = 2$	0.180	0.097	0.233	0.557	0.081	0.154	0.070	0.340	76.7	100
Joinpoint	7.519	0.440	7.982	40.299	0.096	4.327	0.088	11.329	53.7	100

Abbreviations: AIC, Akaike Information Criterion; APC, age-period-cohort; ARIMA, autoregressive integrated moving average; BAPC, Bayesian age-period-cohort; GLM, generalized linear models; LM, linear model. pw, 1/5 power link; RW1, first-order random walk; RW2 second-order random walk.

Largely similar conclusions were obtained in terms of M-IS, with simple models outperforming more complex ones (Table 2 and Figures A5–A8). Again, the best performance was achieved for simple ARIMA models, followed by LM (using the last four points). These simple models also presented an M-CR fairly close to 95%. BAPC models had in general wider prediction intervals, resulting also in too high coverage rates (M-CR close to 100% instead of 95% for some of them). In contrast, coverage rates were clearly too low for GLM and APC, where M-CR could even reach values as low as 50%.

Figure 7 compares the M-NRMSE between model classes separately for each sex and cancer site. Prostate cancer is the site with the largest variability in model accuracy, as incidence rates showed a trend reversal (Figures 4 and A5), followed by skin melanoma for both sexes, because of a rate stabilization in recent years (Figure 2). The APC, BAPC, and GLM models performed particularly poorly in at least one of these difficult cases, while the ARIMA models were among the best methods whatever the cancer site. Here also, similar results were obtained in terms of M-IS (Figure A16).

When evaluating separately the M-NRMSE of the short-term (1–5 years), medium-term (6–10 years), and long-term (11–15 years) predictions, we found that, despite inevitably less good performances for long-term predictions than for mid- and especially short-term predictions, the ranking of the models was largely the same in all three settings (Table 2). Results of our sensitivity analyses (Section 3.2) are also summarized in Table 2. In a first sensitivity analysis, we repeated all calculations using M-NMAE instead of M-NRMSE. The ranking of the models stayed mostly the same. In a second

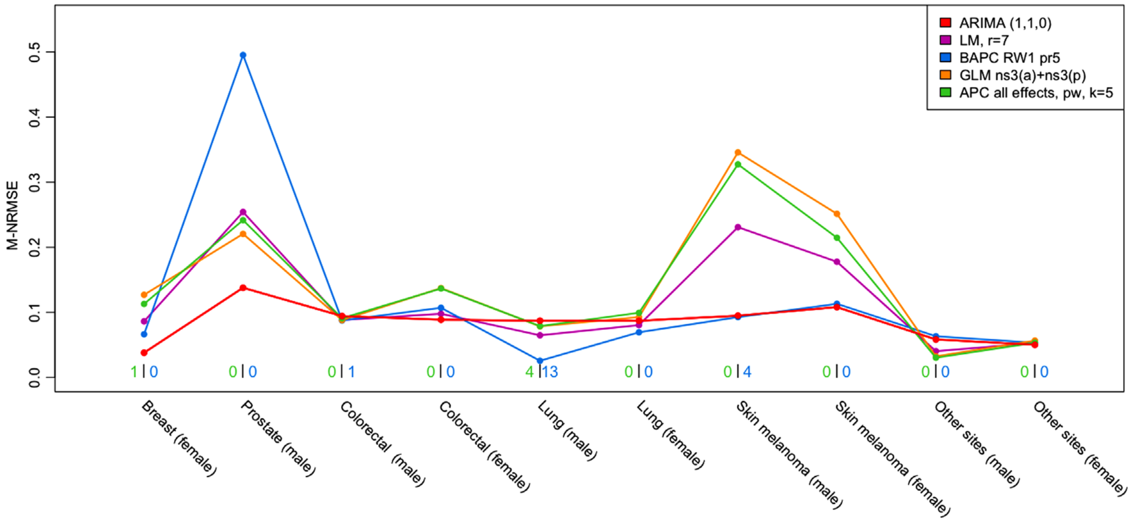

FIGURE 7 Mean normalized root mean square error (M-NRMSE) for one model per class, by sex and cancer site (M-NRSME across 15 leave-future-out cross-validation scenarios on combined data from the Swiss registries of Vaud, Geneva, and Neuchatel, 1982–2016 per sex and cancer site). The number of scenarios for which the fitting algorithm of a method did not converge is indicated at the bottom of the graph for APC|BAPC. The other methods shown always converged. APC, age-period-cohort; ARIMA, autoregressive integrated moving average; BAPC, Bayesian age-period-cohort; GLM, generalized linear models; LM, linear model.

sensitivity analysis, we looked at the Med-NRMSE and Med-NMAE. While the performance of the different models was much closer to each other, the best models remained the simple ARIMA models.

The last column of Table 2 provides the percentage of scenarios for which the fitting algorithm of the different methods converged, which was 100% for LM, above 95% for ARIMA and GLM, and slightly below (down to 85%) for the APC and BAPC models. The computational time of Bayesian and non-Bayesian methods is available in the Supporting Information Appendix (Table A1).

6 | DISCUSSION

The objective of this paper was to compare the performance of statistical models for predicting annual age-standardized cancer incidence rates. The APC method was compared with its Bayesian counterpart BAPC, with simpler GLM approaches including only age and period or age and cohort in the model, and with ARIMA models based only on time series of incidence rates. A simple linear regression model (LM) extrapolating the trend fitted on the last few observed incidence rates was included in the comparison. The prediction performance of each method was evaluated in terms of point prediction accuracy, the quality of prediction intervals, and the convergence rate of the fitting algorithm on a large set of scenarios obtained by leave-future-out cross-validation using real data from Swiss cancer registries.

In the class of GLM, including period led to better prediction performance than including cohort (in addition to age) in the model. Including cohort in the model involves the prediction of new cohort effects mainly based on the most recent ones containing only few observations. These models are therefore more prone to overfitting than those using period effects concerning the whole population. On the other hand, the inclusion of a simple (linear) interaction term for age and period did not improve prediction performance, while including complex interactions (splines) actually worsened predictions due to overfitting. This was also the case for the Lee–Carter and joinpoint models, and for a GLM model selected via the AIC. Prediction intervals were particularly poor for GLM as previously found by another study (Møller et al., 2005), who estimated (as we did) coverage rates of about 50%. The APC approach, although considerably improved by the use of a power link (Møller et al., 2003), performed poorly when predictions were made by extrapolating the drift alone. The drift alone is, however, the most widely used strategy to date (Yu et al., 2019). Further extrapolating all nonlinear effects (Rutherford et al., 2012) improved the predictions in case of a trend reversal. The Bayesian APC models based on second-order random walk performed very similarly to the classical APC models with splines, while a simpler BAPC based on first-order random walk performed slightly better despite less smooth age, period, and cohort effects. For ARIMA, models with low-order autoregressive, integrated, and moving average components are to be preferred to more complex model structures. A strategy based on selecting the best ARIMA model via AIC overfitted the data and gave worse predictions than systematically choosing a simple structure. Finally, increasing the number of data points to fit a simple linear regression model improved the prediction when including up to 7 years in the past, worsening after this lag.

Comparing models among classes, we obtained the best performance for simple ARIMA models such as (2,1,1) or (1,1,0). Nevertheless, most of the simplest ARIMA models, including component orders up to (3,1,3), were ahead of the other methods. The second-best performing class of models was simple linear regression, where all alternatives using 5–10 data points to fit the trend outperformed the more complex GLM, APC, and BAPC models. Similar results in terms of ranking of models were obtained for short, medium, and long-term predictions, and stratifying by cancer site, despite the less good performance for long-term versus short-term predictions, and for cancers showing a trend reversal. Considering absolute errors (NMAE) instead of square errors (NRMSE) and taking the median instead of the mean of the NRMSE (or NMAE) across scenarios did not substantially change the ranking of the models, still placing simple ARIMA models at the top of the list. However, while the choice of the mean strongly penalizes the most difficult scenarios, the median simply excludes these particularly complicated situations from the evaluation. For this reason, we have focused mainly on the mean NRMSE, as it allows us to favor models and methods that do not show aberrant behavior in case of trend reversal or stabilization, as this is becoming increasingly common (e.g., prostate, breast, and skin melanoma) and a similar pattern will hopefully be observed in the future for other cancer sites. Finally, looking at the prediction intervals did not alter our conclusions, with simple ARIMA models still being the best performing models.

Our first observation is that the models that offer the best fit to the data, for example, models selected using AIC, are not necessarily the best for prediction. The use of AIC may lead to the choice of overly complex models, for example, GLM or APC including complex interactions with splines, which are likely to overfit the data and are indeed less efficient for prediction than simpler and smoother models. The second observation is in line with the first one by advocating for simplicity. Approaches such as APC or BAPC estimate the effects of age, period, and/or cohort separately, considered as proxies for epidemiological risk factors, such as smoking and other carcinogens. However, when it is a matter of prediction, we have shown that better performance is obtained by simply extrapolating the trends using models such as ARIMA (or even simple regression models). As Booth and Tickle (2008) have already noted in the context of mortality projection, an extrapolation-based approach is often preferable to an explanation- or interpretation-based approach.

Our study has some limitations, all of which could motivate future work. For example, although we considered many models, we did not include overdispersion in our Poisson models. While this would affect the coverage of the prediction intervals, it would not greatly change the predictions and thus the NRMSE, our main criterion. More generally, many other prediction models and options can be formulated, for example, using quasi-Poisson or negative binomial distributions or trying other link functions, and it could become an interesting challenge for researchers to try to identify one that can outperform ARIMA in comparable settings. Second, we did not take into account the effects of prevention/screening programs and changes in medical technology and practice on cancer incidence rates. These factors can strongly influence the (past and future) rates and help identify trend changes (Etzioni et al., 2013). Such effects are however challenging to analyze, as they evolve over time. For example, the start of a cancer screening program leads to a short-term initial increase in incidence, followed by a stabilization or sometimes a decrease compared to past trends. The magnitude of these temporal effects will depend on the adherence to these prevention programs. Considering the impact of screening when making predictions is a subject of ongoing work. Another limitation is that we did not account for data underreporting and its evolution over time, which may introduce artifacts into predictions. In fact, while a high degree of completeness has been evaluated in Swiss cancer registration (Lorez et al., 2017), a slight underreporting of some cancers cannot be discarded. Finally, we have only modeled the most common cancers. Although we assume that simpler methods, such as ARIMA, should be suitable for rare cancers, as they are less prone to overfitting, we have not studied the prediction performance for these cancers and cannot draw any conclusions on this point. A specific analysis of the most suitable prediction methods for rare cancers would be another important subject of future work. The same consideration can be made for the prediction of specific age incidence rates.

In conclusion, we recommend using lower order ARIMA models to predict cancer incidence, for example, the ARIMA (2,1,1) or (1,1,0) models, which achieved the best overall performance in our comparison study. We suggest not using AIC to select the model, as this appears not to be the best strategy for prediction because it often results in overfitting. While APC and BAPC models remain the best to help interpret changes in cancer incidence trends, we recommend avoiding the widely used APC model for prediction, especially when the extrapolation is restricted to the drift. Finally, given the large uncertainty, particularly in the case of a trend reversal, we do not recommend predicting cancer incidence for periods far into the future but rather updating predictions regularly.

ACKNOWLEDGMENT

This work has been supported by the Krebsliga Schweiz (KFS-4385-02-2018) (Isabella Locatelli).

Open access funding provided by Universite de Lausanne.

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

Population data are available on the FSO website (https://www.pxweb.bfs.admin.ch/pxweb/fr/px-x-0102020000_104/px-x-0102020000_104/px-x-0102020000_104.px) and cancer incidence data can be requested from cancer registries after ethical approval of a project. In the Supporting Information, a simulated dataset, close to the original one is available, as well as the R code used to produce the analysis, tables, and figures presented in the article.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data confidentiality issues.

ORCID

Bastien Trächsel https://orcid.org/0000-0002-4519-094X
Jean-Luc Bulliard https://orcid.org/0000-0001-9750-2709

REFERENCES

Booth, H., & Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, *3*(1–2), 3–43. https://doi.org/10.1017/S1748499500000440

Bürkner, P.-C., Gabry, J., & Vehtari, A. (2020). Approximate leave-future-out cross-validation for Bayesian time series models. *Journal of Statistical Computation and Simulation*, *90*(14), 2499–2523. https://doi.org/10.1080/00949655.2020.1783262

Cameron, A. C., & Trivedi, P. K. (2013a). 2.4.4 Generalized linear models. In *Regression analysis of count data* (2nd ed., pp. 35–38). Cambridge University Press. https://doi.org/10.1017/CBO9781139013567

Cameron, A. C., & Trivedi, P. K. (2013b). 3 Basic count regression. In *Regression Analysis of Count Data* (2nd ed., p. 566). Cambridge University Press. https://doi.org/10.1017/CBO9781139013567

Cameron, J. K., & Baade, P. (2021). Projections of the future burden of cancer in Australia using Bayesian age-period-cohort models. *Cancer Epidemiology*, *72*, 101935. https://doi.org/10.1016/j.canep.2021.101935

Carstensen, B. (2007). Age–period–cohort models for the Lexis diagram. *Statistics in Medicine*, *26*(15), 3018–3045. https://doi.org/10.1002/sim.2764

Carstensen, B., Plummer, M., Laara, E., & Hills, M. (2022). *Epi: A package for statistical analysis in epidemiology*. https://CRAN.R-project.org/package=Epi

Chen, H. S., Portier, K., Ghosh, K., Naishadham, D., Kim, H.-J., Zhu, L., Pickle, L. W., Krapcho, M., Scoppa, S., Jemal, A., & Feuer, E. J. (2012). Predicting US and state-level cancer counts for the current calendar year: Part I: Evaluation of temporal projection methods for mortality. *Cancer*, *118*(4), 1091–1099. https://doi.org/10.1002/cncr.27404

Chen, W.-Q., Zheng, R.-S., & Zeng, H.-M. (2011). Bayesian age-period-cohort prediction of lung cancer incidence in China: Predicted lung cancer incidence in China. *Thoracic Cancer*, *2*(4), 149–155. https://doi.org/10.1111/j.1759-7714.2011.00062.x

Clayton, D., & Schifflers, E. (1987). Models for temporal variation in cancer rates. II: Age–period–cohort models. *Statistics in Medicine*, *6*(4), 469–481. https://doi.org/10.1002/sim.4780060406

Clements, M. S. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics*, *6*(4), 576–589. https://doi.org/10.1093/biostatistics/kxi028

Coupland, V. H., Okello, C., Davies, E. A., Bray, F., & Moller, H. (2010). The future burden of cancer in London compared with England. *Journal of Public Health*, *32*(1), 83–89. https://doi.org/10.1093/pubmed/fdp082

Dunn, P. K., & Smyth, G. K. (2018). Chapter 12: Tweedie GLMs. In P. K. Dunn & G. K. Smyth, *Generalized linear models with examples in R* (pp. 457–490). Springer. https://doi.org/10.1007/978-1-4419-0118-7_12

Earnest, A., Evans, S. M., Sampurno, F., & Millar, J. (2019). Forecasting annual incidence and mortality rate for prostate cancer in Australia until 2022 using autoregressive integrated moving average (ARIMA) models. *BMJ Open*, *9*(8), e031331. https://doi.org/10.1136/bmjopen-2019-031331

Etzioni, R., Gulati, R., Mallinger, L., & Mandelblatt, J. (2013). Influence of study features and methods on overdiagnosis estimates in breast and prostate cancer screening. *Annals of Internal Medicine*, *158*(11), 831. https://doi.org/10.7326/0003-4819-158-11-201306040-00008

Fritz, A., Percy, C., Jack, A., Shanmugaratnam, K., Sobin, L. H., Parkin, D. M., Whelan, S. L., & World Health Organization. (2000). *International classification of diseases for oncology*. [Classification internationale des maladies pour l'oncologie]. WHO IRIS. https://apps.who.int/iris/handle/10665/42344

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691042893/time-series-analysis

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. https://doi.org/10.1007/978-0-387-84858-7

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, *39*(2), 311–324. https://doi.org/10.2307/2531004

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, *87*(419), 659–671. https://doi.org/10.1080/01621459.1992.10475265

Lerman, P. M. (1980). Fitting segmented regression models by grid search. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(1), 77–84. https://doi.org/10.2307/2346413

Li, C., Zhu, Y., Yang, J., Xu, D., Wang, J., Chen, K., & Li, Q. (2022). Incidence of lung cancer in Jiashan, Zhejiang province: Trend analysis from 1987 to 2016 and projection from 2017 to 2019. *Journal of Zhejiang University (Medical Sciences)*, *47*(4), 367–373. https://doi.org/10.3785/j.issn.1008-9292.2018.08.07

Lin, H., Shi, L., Zhang, J., Zhang, J., & Zhang, C. (2021). Epidemiological characteristics and forecasting incidence for patients with breast cancer in Shantou, Southern China: 2006–2017. *Cancer Medicine*, *10*(8), 2904–2913. https://doi.org/10.1002/cam4.3843

Lindgren, F., & Rue, H. (2008). On the second-order random walk model for irregular locations: RW2 for irregular locations. *Scandinavian Journal of Statistics*, *35*(4), 691–700. https://doi.org/10.1111/j.1467-9469.2008.00610.x

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25. https://doi.org/10.18637/jss.v063.i19

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Lorez, M., Bordoni, A., Bouchardy, C., Bulliard, J.-L., Camey, B., Dehler, S., Frick, H., Konzelmann, I., Maspoli, M., Mousavi, S. M., Rohrmann, S., & Arndt, V. (2017). Evaluation of completeness of case ascertainment in Swiss cancer registration. *European Journal of Cancer Prevention*, *26*, S139–S146. https://doi.org/10.1097/CEJ.0000000000000380

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, *67*, 68–83. https://doi.org/10.1016/j.csda.2013.04.014

Møller, B., Fekjaer, H., Hakulinen, T., Sigvaldason, H., Storm, H. H., Talbäck, M., & Haldorsen, T. (2003). Prediction of cancer incidence in the Nordic countries: Empirical comparison of different approaches: Comparison of methods for incidence prediction. *Statistics in Medicine*, *22*(17), 2751–2766. https://doi.org/10.1002/sim.1481

Møller, B., Fekjaer, H., Hakulinen, T., Tryggvadóttir, L., Storm, H. H., Talbäck, M., & Haldorsen, T. (2002). Prediction of cancer incidence in the Nordic countries up to the year 2020. *European Journal of Cancer Prevention*, *11*(Suppl 1), S1–96.

Møller, B., Weedon-Fekjær, H., & Haldorsen, T. (2005). Empirical evaluation of prediction intervals for cancer incidence. *BMC Medical Research Methodology*, *5*(1), 21. https://doi.org/10.1186/1471-2288-5-21

NSW Cancer Institute. (2016). *Cancer incidence and projections 2011–2021*. Cancer Institute.

OFS. (2020). *Cancer, nouveaux cas et décès: Nombre, taux et évolution par localisation cancéreuse et période - 1987–2017 | Tableau*. Office fédéral de la statistique. https://www.bfs.admin.ch/asset/fr/14816247

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rahib, L., Wehner, M. R., Matrisian, L. M., & Nead, K. T. (2021). Estimated projection of US cancer incidence and death to 2040. *JAMA Network Open*, *4*(4), e214708. https://doi.org/10.1001/jamanetworkopen.2021.4708

Rapiti, E., Guarnori, S., Pastoors, B., Miralbell, R., & Usel, M. (2014). Planning for the future: Cancer incidence projections in Switzerland up to 2019. *BMC Public Health*, *14*(1), 102. https://doi.org/10.1186/1471-2458-14-102

Riebler, A., & Held, L. (2017). Projecting the future burden of cancer: Bayesian age-period-cohort analysis with integrated nested Laplace approximations: Projecting the future burden of cancer. *Biometrical Journal*, *59*(3), 531–549. https://doi.org/10.1002/bimj.201500263

Riebler, A., Held, L., & Rue, H. (2012). Estimation and extrapolation of time trends in registry data—Borrowing strength from related populations. *The Annals of Applied Statistics*, *6*(1), 304–333. https://doi.org/10.1214/11-AOAS498

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(2), 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511755453

Rutherford, M. J., Thompson, J. R., & Lambert, P. C. (2012). Projecting cancer incidence using age-period-cohort models incorporating restricted cubic splines. *The International Journal of Biostatistics*, *8*(1),33. https://doi.org/10.1515/1557-4679.1411

Schmid, V., & Held, L. (2004). Bayesian extrapolation of space-time trends in cancer registry data. *Biometrics*, *60*(4), 1034–1042. JSTOR. https://doi.org/10.1111/j.0006-341X.2004.00259.x

Schumacher, D., & Jombart, T. (2021). *trending: Model temporal trends*. https://CRAN.R-project.org/package=trending

Shi, J., Cao, M., Wang, Y., Bai, F., Lei, L., Peng, J., Feletto, E., Canfell, K., Qu, C., & Chen, W. (2021). Is it possible to halve the incidence of liver cancer in China by 2050? *International Journal of Cancer*, *148*(5), 1051–1065. https://doi.org/10.1002/ijc.33313

Spiegelman, M., & Marks, H. H. (1966). Empirical testing of standards for the age adjustment of death rates by the direct method. *Human Biology*, *38*(3), 279–292.

Tsoi, K. K. F., Hirai, H. W., Chan, F. C. H., Griffiths, S., & Sung, J. J. Y. (2017). Cancer burden with ageing population in urban regions in China: Projection on cancer registry data from World Health Organization. *British Medical Bulletin*, *121*(1), 83–94. https://doi.org/10.1093/bmb/ldw050

Wong, M. C. S., Huang, J., Lok, V., Wang, J., Fung, F., Ding, H., & Zheng, Z.-J. (2021). Differences in incidence and mortality trends of colorectal cancer worldwide based on sex, age, and anatomic location. *Clinical Gastroenterology and Hepatology*, *19*(5), 955–966. e61. https://doi.org/10.1016/j.cgh.2020.02.026

Yang, L., Parkin, D. M., Ferlay, J., Li, L., & Chen, Y. (2005). Estimates of cancer incidence in China for 2000 and projections for 2005. *Cancer Epidemiology, Biomarkers & Prevention*, *14*(1), 243–250. https://doi.org/10.1158/1055-9965.243.14.1

Yu, X. Q., Luo, Q., Hughes, S., Wade, S., Caruana, M., Canfell, K., & O'Connell, D. L. (2019). Statistical projection methods for lung cancer incidence and mortality: A systematic review. *BMJ Open*, *9*(8), e028497. https://doi.org/10.1136/bmjopen-2018-028497

Zemni, I., Kacem, M., Dhouib, W., Bennasrallah, C., Hadhri, R., Abroug, H., Ben Fredj, M., Mokni, M., Bouanene, I., & Belguith, A. S. (2022). Breast cancer incidence and predictions (Monastir, Tunisia: 2002–2030): A registry-based study. *PLoS ONE*, *17*(5), e0268035. https://doi.org/10.1371/journal.pone.0268035

Zheng, Y., Zhang, L., Zhu, X., & Guo, G. (2020). A comparative study of two methods to predict the incidence of hepatitis B in Guangxi, China. *PLoS ONE*, *15*(6), e0234660. https://doi.org/10.1371/journal.pone.0234660

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Trächsel, B., Rousson, V., Bulliard, J.-L., & Locatelli, I. (2023). Comparison of statistical models to predict age-standardized cancer incidence in Switzerland. *Biometrical Journal*, e2200046. https://doi.org/10.1002/bimj.202200046