

## RESEARCH ARTICLE

# Using mortality to predict incidence for rare and lethal cancers in very small areas

Jaione Etxeberria<sup>1,2</sup> | Tomás Goicoa<sup>1,2,3</sup> | Maria D. Ugarte<sup>1,2</sup> 

<sup>1</sup>Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Campus Arrosadia, Pamplona, Navarre, Spain

<sup>2</sup>Institute for Advanced Materials and Mathematics (INAMAT<sup>2</sup>), Public University of Navarre (UPNA), Campus Arrosadia, Pamplona, Navarre, Spain

<sup>3</sup>Research Network on Health Services in Chronic Diseases (REDISSEC), Madrid, Spain

## Correspondence

M.D. Ugarte, Department of Statistics, Computer Science and Mathematics, Public University of Navarre (UPNA), Campus Arrosadia, 31006 Pamplona, Navarre, Spain.

Email: [lola@unavarra.es](mailto:lola@unavarra.es)

## Funding information

Spanish Research Agency, Grant/Award Number: PID2020-113125RB-I00/MCIN/AEI/10.13039/501100011033; Universidad Pública de Navarra, Grant/Award Number: PJUPNA2018-11



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

Incidence and mortality figures are needed to get a comprehensive overview of cancer burden. In many countries, cancer mortality figures are routinely recorded by statistical offices, whereas incidence depends on regional cancer registries. However, due to the complexity of updating cancer registries, incidence numbers become available 3 or 4 years later than mortality figures. It is, therefore, necessary to develop reliable procedures to predict cancer incidence at least until the period when mortality data are available. Most of the methods proposed in the literature are designed to predict total cancer (except nonmelanoma skin cancer) or major cancer sites. However, less frequent lethal cancers, such as brain cancer, are generally excluded from predictions because the scarce number of cases makes it difficult to use univariate models. Our proposal comes to fill this gap and consists of modeling jointly incidence and mortality data using spatio-temporal models with spatial and age shared components. This approach allows for predicting lethal cancers improving the performance of individual models when data are scarce by taking advantage of the high correlation between incidence and mortality. A fully Bayesian approach based on integrated nested Laplace approximations is considered for model fitting and inference. A validation process is also conducted to assess the performance of alternative models. We use the new proposals to predict brain cancer incidence rates by gender and age groups in the health units of Navarre and Basque Country (Spain) during the period 2005–2008.

## KEYWORDS

brain cancer incidence, disease mapping, INLA, predictions, shared component models

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

## 1 | INTRODUCTION

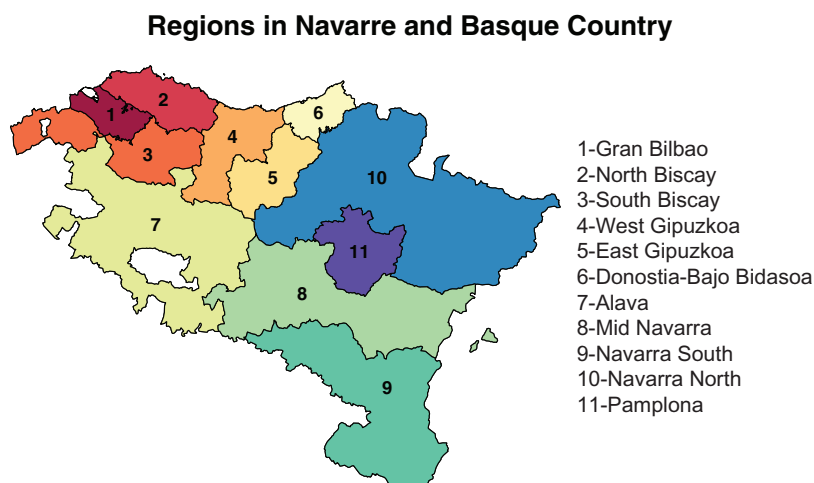
Cancer incidence predictions play an important role in epidemiology allowing cancer monitoring in a population even in the absence of specific control plans. For administrative purposes, predictions are also useful to support public health decision-making processes related to interventions, screening, cancer control programs, treatments, and rehabilitation. Cancer predictions have different purposes. In countries without a national cancer registry, the interest resides in estimating or predicting cancer incidence at the national level. For that aim, different prediction techniques have been developed by Galceran et al. (2017), Uhry et al. (2007), and Ferlay et al. (2018). In some countries, such as Spain, France, or Italy, cancer incidence figures are monitored by provincial cancer registries covering only a part of the population (see Figure A.1 in the Appendix); nevertheless, mortality numbers are provided by National Statistical Offices making them available at different levels (municipality, province, autonomous region, or national level). In this context, different approaches and statistical models have been developed to estimate national incidence using national mortality and the polled incidence data provided by the local registries (Møller et al., 2003).

Due to the complexity of updating cancer registries, incidence numbers become available 3 or 4 years later than mortality figures. The completeness of a cancer registry database is a very important quality requirement. Therefore, cancer registries, researchers, and local health decision-makers are highly interested in developing reliable procedures to complete the registry databases at least up to the period when the mortality data are available. Most of the methods proposed in the literature stem from the recommendations of the International Agency for Research on Cancer (IARC) on how realistic predictions should be done. According to this agency, predictions of cancer incidence should fulfill a list of requirements (Bray et al., 2013). First, predictions should be smooth over time. Abrupt changes in time trends may lead to the appearance of unexpected or implausible incidence trends within a registry's dataset. Second, they must be comparable in different populations or regions. This allows to identify high or low incidence patterns by specific regions. Third, age-specific incidence curves should be provided, including childhood cancer rates. Incidence rates of cancer in children tend to be lower than the rates in adults, although there are some well-documented geographical and ethnic differences for certain pediatric cancers, such as Brain and Central Nervous System cancer (hereafter BCNS) or leukemia (Steliarova-Foucher et al., 2017). Unexpected drops in age-specific trends may indicate problems with source files, for example, the size of the populations at risk in the age groups.

Finally, the mortality-to-incidence (M/I) ratio should be taken into account. This ratio compares the number of deaths due to a specific type of cancer over a specific period of time (usually obtained from a source that is independent of the registry such as National Statistical Offices) with the number of new cases of that type of cancer registered during the same period by the cancer registry. This ratio is also an important indicator of completeness as long as the quality of the mortality data is good. Usually, the observed M/I ratios for a specific registry are compared to the values obtained for a similar cancer registry or region. M/I ratios higher than expected raise suspicions of incompleteness.

Based on all these recommendations, different methods have been proposed in the literature. The very first procedures come from the Finnish Cancer Registry (Hakulinen et al., 1986; Teppo et al., 1974), and they are based on the linear extrapolation of cancer incidence trends. However, age-period-cohort (APC) models (Holford, 1983; Osmond, 1985) have been historically the most popular tools (Dyba & Hakulinen, 2000; Møller et al., 2003). Different versions of the APC models were developed by Møller et al. (2003). In particular, different link functions between rates and covariates were employed (the log link and the power link), and shorter and longer observed time trends were used. At a local or national level, research has been conducted to predict cancer incidence rates based on the previous methodologies. Most of the literature provides incidence and mortality estimates and predictions for total cancer and/or for the most common cancer types such as breast, prostate, or colorectal cancer (Bezerra-de Souza et al., 2012; Sánchez et al., 2010). Less frequent cancer sites such as brain, pancreatic, or ovarian cancer are generally excluded from predictions. The main reason to do this is because the aforementioned APC models require a disaggregation of the number of cases by age group and calendar year. However, data scarcity leads to imprecise incidence forecasts when these methods are used and, therefore, predicting rare or less frequent cancers becomes a challenge from a methodological point of view. As far as we know, there is no specific methodology to solve this problem and we therefore propose a joint modeling method with spatial and age-shared components that elegantly exploits the correlation between cancer incidence and mortality to improve incidence forecasts of rare cancer types. Here we illustrate the methodology by predicting BCNS incidence rates in subregions of Navarra and Basque Country, two northern regions of Spain that have historically presented very high BCNS incidence rates compared to other regions in Europe (Ferlay et al., 2013). This cancer is very lethal with a high correlation between incidence and mortality. Hence, it is the necessity of careful monitoring over time. Our approach takes into consideration the previ-

**FIGURE 1** Navarre (regions 8–11 on the right) and the Basque Country (Regions 1–7 on the left), Spain



ous recommendation of the IARC as different age, time, gender, and spatial-specific terms are considered in the models. Moreover, our proposal is an interesting strategy to predict incidence for rare and lethal cancers because the multivariate modeling smartly overcomes sparsity by putting together two sources of information, mortality, and incidence.

The rest of the paper is laid out as follows. In Section 2, an exploratory data analysis is provided to set the problem. Section 3 describes a set of joint models predicting cancer incidence, and how computation, model parameter estimation, and prediction are conducted. A validation process is presented in Section 4. Results are shown in Section 5. Finally, the paper ends with a discussion.

## 2 | BCNS INCIDENCE AND MORTALITY DATA FROM NORTHERN SPAIN

Navarre and the Basque Country are two regions located in northern Spain ranked among the European regions with the highest rates of BCNS (Forman et al., 2013). More precisely, Navarre and Basque Country are in the ninth and 19th position out of 119 in the ranking of regions with the highest rates (both genders) in Europe. Previous geographical analysis in Spain also showed a cluster of high risk in these regions. Some of these investigations were motivated by the possible association between BCNS and the types of soil cover and/or crop and plant protection treatments used in rural areas, but no evidence was found. Despite the efforts to identify BCNS risk factors, very little progress has been made. Besides exposure to ionizing radiation, no other definitive risk factor is known (Connelly & Malkin, 2007; Ugarte et al., 2015a).

Our study is based on incidence cases and deaths of brain and central nervous system tumors (C70–C72, International Classification of Diseases-10) reported by the regional population-based cancer registries of Navarre and the Basque Country. Data are organized by age group, gender, period, and region. More precisely, data are split by 18 age groups, gender, regions, and calendar year (1989–2008 for mortality and 1989–2004 for incidence). Figure 1 displays the regions of Navarre and the Basque Country considered in this paper. The regions are numbered from 1 to 11. Regions 1–7 belong to the Basque Country (1–3 to the province of Vizcaya, regions 4–6 to the province of Gipuzkoa, and region 7 represents the province of Alava). Finally, regions 8–11 belong to the province of Navarre.

A total of 3615 cases of malignant brain tumors between 1989 and 2004 (55.29% males and 44.71% females) and 3296 deaths between 1989 and 2008 (55.10% males and 44.90% females) were reported by the two cancer registries, representing on average 225 incidence and 165 mortality cases per year. Crude incidence and mortality rates of brain cancer per 100,000 inhabitants were calculated using 18 age groups, the two genders, and all the regions. Similar overall crude incidence and mortality rates were observed (6.8 and 6.20 cases per 100,000 inhabitants, respectively).

Figure 2 shows age-specific incidence (continuous line) and mortality (dashed line) rates for males (blue) and females (red), respectively, during the study period. Although the distribution is similar in shape in both sexes, differences can be observed with males having higher incidence and mortality rates in all age groups. Both incidence and mortality rates peak in the 65–80 age group, decreasing for 80+. There is also a small peak in incidence rates in early childhood (0–4, and 5–9 age groups) in both sexes. This is not very common in other cancer sites. This exploratory analysis shows that gender and age at diagnosis are particularly important in characterizing brain cancer. Similar to other research on rare cancer types (Etxeberria et al., 2017) and to ensure a sufficient number of cases to allow model fitting and prediction, the

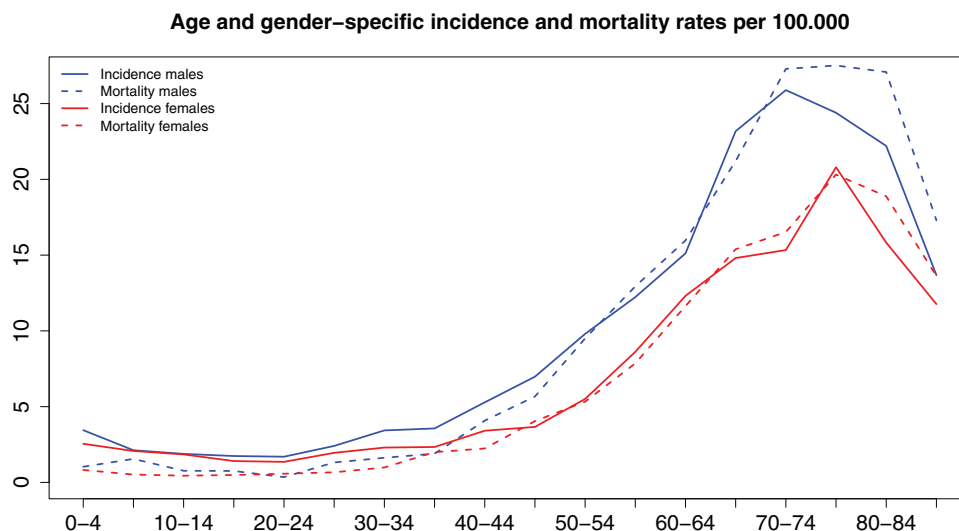


FIGURE 2 Age and gender-specific incidence and mortality rates during the study period

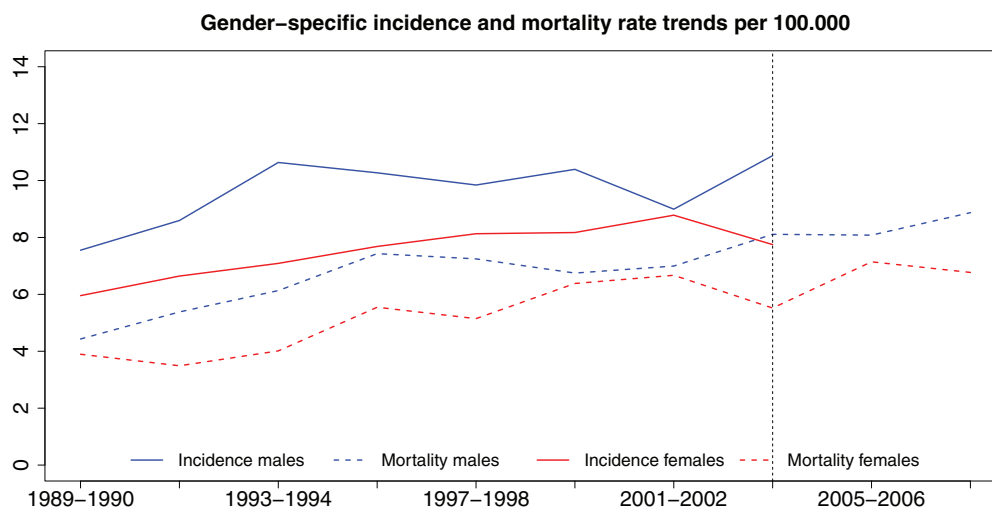


FIGURE 3 Crude incidence and mortality rates trends by gender

18 groups are now reorganized into the following age groups <40, 40-49, 50-59, 60-69, 70-79, and 80+, and the period is managed on a biannual basis, 1989-1990, 1991-1992, ..., 2007-2008. Data scarcity and the consequent huge variability preclude the analysis if yearly data are considered.

Male (blue) and female (red) global trends of crude incidence and mortality rates are depicted in Figure 3 from 1989 to 2008. Note that incidence is only considered up to 2004. In males, the crude incidence rates increase up to 1994, they decrease up to 2000 and experience a V-shaped trend up to 2004. Incidence rates for females present an increasing trend up to 2002 and a slight decrease in the past 2 years. Crude mortality rates show an upward trend throughout the entire period for both genders.

Top panels in Figure 4 display crude incidence (left) and mortality rates (right) per 100,000 inhabitants by region. In this figure, regions located in the north and mid-Navarre are the ones presenting the highest incidence and mortality rates. Overall, the geographical patterns of incidence and mortality are not very different, suggesting a high correlation between them. This is confirmed by the scatter plot of incidence and mortality rates by region at the bottom panel of Figure 4.

The exploratory data analysis provides a preliminary idea of how brain cancer incidence and mortality behave by age group, gender, region, and time. This information is very useful to define suitable models that can appropriately fit the data.

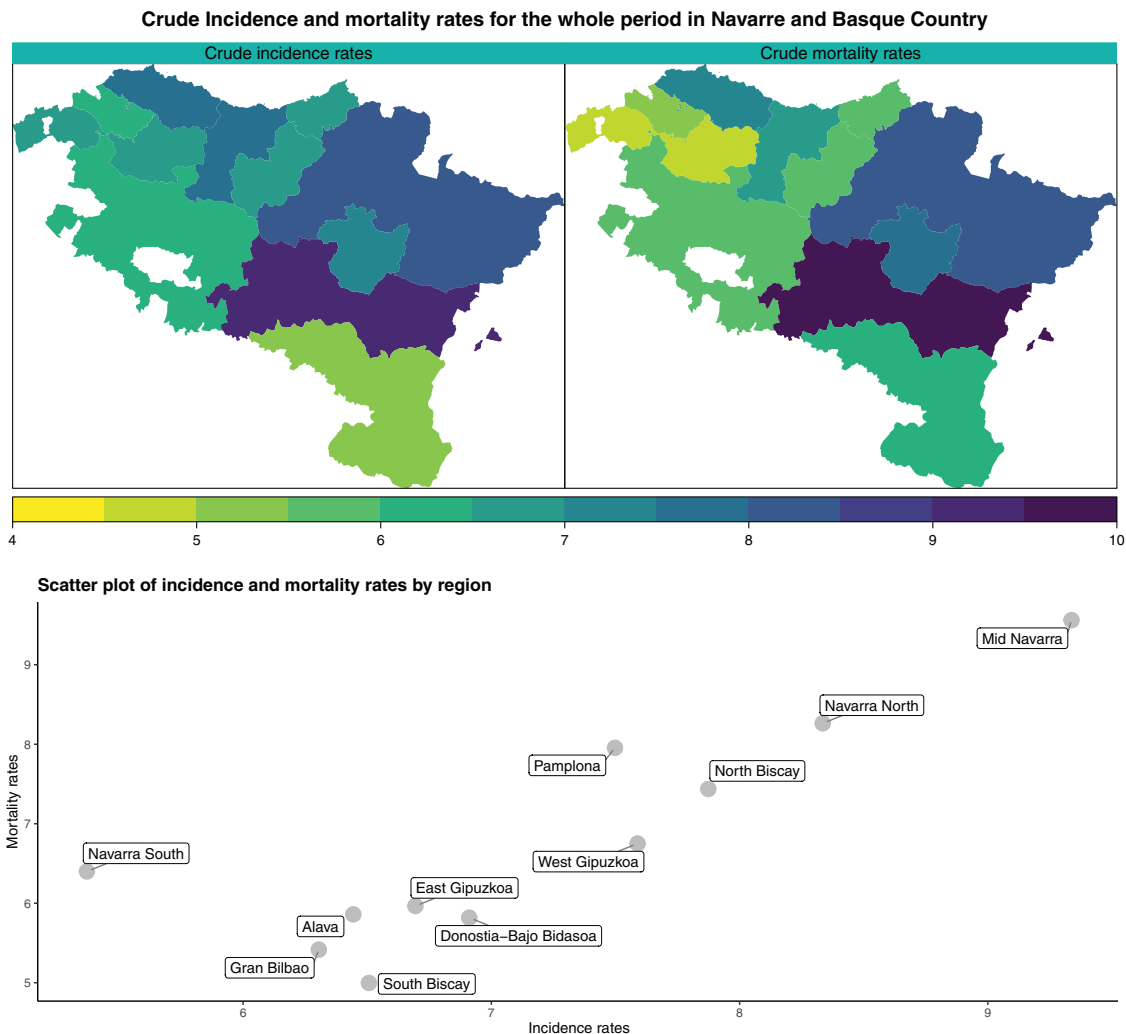


FIGURE 4 Crude incidence and mortality rates by region for both genders (top panels) and scatter plot of incidence and mortality rates by region (bottom panel)

### 3 | MODELS TO PREDICT CANCER INCIDENCE USING MORTALITY DATA

In this section, different age and gender-specific shared component models are proposed to predict cancer incidence. These shared component models constitute a simple way of modeling several diseases, and they can be embedded within the general multivariate framework (MacNab, 2010). For a general review of these types of models in disease mapping, the reader is referred to some recent work by MacNab (2016a, 2016b). One difference between shared component models (SCM) and more general multivariate models is that in SCMs, dependence between diseases is assumed a priori whereas multivariate models are more appropriate if the relationship among diseases is unknown. Here we exploit the correlation between incidence and mortality in BNCS, and hence we propose shared component models.

The context of our study is the following. Let us define as  $O_{1igt}$  and  $O_{2igt}$  the number of incidence and mortality cases, respectively by health-area  $i = 1, \dots, n = 11$ , gender  $g$  (male or female), age group  $j = < 40, 40-49, 50-59, 60-69, 70-79$ , and  $80+$  and, time period  $t = 1, \dots, 10$  where  $1 = 1989-1990, 2 = 1991-1992, \dots, 10 = 2007-2008$ . Incidence data,  $O_{1igt}$ , are only available for  $t = 1, \dots, 8$ . In the first level of the Bayesian hierarchical structure, the likelihood, we assume that conditional on the rates,  $O_{1igt}$  and  $O_{2igt}$  follow the next Poisson distributions

$$O_{1igt} | r_{1igt} \sim \text{Poisson}(\mu_{1igt} = n_{igt} r_{1igt}), \quad \log \mu_{1igt} = \log n_{igt} + \log r_{1igt}, \tag{1}$$

$$O_{2igt} | r_{2igt} \sim \text{Poisson}(\mu_{2igt} = n_{igt} r_{2igt}), \quad \log \mu_{2igt} = \log n_{igt} + \log r_{2igt}. \tag{1}$$

In these expressions,  $n_{igt}$  is the population at risk (the same for incidence and mortality) and  $r_{1igt}$  and  $r_{2igt}$  are the incidence and mortality rates in region  $i$ , gender  $g$ , age group  $j$ , and period  $t$ . Recall that for  $t = 9, 10$  and all  $i, g, j$ , the observed incidence rates are unavailable.

The interest here relies on modeling the log incidence rates ( $\log r_{1igt}$ ) and log mortality rates ( $\log r_{2igt}$ ) jointly and, therefore, to get an appropriate posterior predictive distribution for the nonobserved incidence cases. For this aim, a set of models are proposed. For ease of reading, only some of them are described in this paper. Due to the important role that gender and age groups play in describing brain cancer incidence and mortality patterns, we consider models incorporating space, time, age group, gender, and interactions between them. Let us first start with model 1 (M1) including a gender-specific shared component spatial term, and age and time effects common to both incidence and mortality.

$$\text{M1} : \log r_{1igt} = \delta_g \phi_{ig} + \gamma_t + \xi_j, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \gamma_t + \xi_j. \quad (2)$$

In these expressions,  $\phi_{\text{males}} = (\phi_{1m}, \dots, \phi_{nm})'$  and  $\phi_{\text{females}} = (\phi_{1f}, \dots, \phi_{nf})'$  are assumed to follow multivariate normal distributions. Namely,  $\phi_{\text{males}} \sim N(\mathbf{0}, \sigma_{\phi_{\text{males}}}^2 \mathbf{Q}^-)$  and  $\phi_{\text{females}} \sim N(\mathbf{0}, \sigma_{\phi_{\text{females}}}^2 \mathbf{Q}^-)$ , respectively, where  $\mathbf{Q}$  is the spatial neighborhood matrix defined by Besag et al. (1991). The symbol  $^-$  denotes the Moore–Penrose generalized inverse. Here two areas are considered neighbors if they share a common border. Note that the simplest shared component model defined by Knorr-Held and Best (2001), includes an additional parameter  $\delta$ , where  $\delta^2$  can be interpreted as the ratio between log-incidence and log-mortality gradients. Prediction models including shared component terms are appropriate as they allow to monitor how much of the spatial pattern is common to both mortality and incidence, how much is specific to each one, and to interpret  $1/\delta^2$  as a kind of mortality to incidence ratio, something recommended by the IARC. Moreover, in this work models including gender-specific parameters  $\delta_g = (\delta_{\text{males}}, \delta_{\text{females}})$  are considered. This idea comes from the work by Etxeberria et al. (2018) in which different spatial shared component models are examined. In particular, they compare gender-specific shared spatial components in which the same parameter  $\delta$  or gender-specific parameters  $\delta_g$  are considered. Introducing gender-specific parameters makes the model more flexible, and it provides better results, as the spatial component is allowed to be different between genders with the same or different precision parameters controlling the degree of smoothing. Additionally, a common temporal random effect and another common age effect for incidence and mortality are considered in model 1 (M1) assuming the following distributions for the vectors  $\gamma$  and  $\xi$ :

- $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_{10})' \sim N(\mathbf{0}, \sigma_{\gamma}^2 \mathbf{Q}_T^-)$  represents the time-specific random effect and
- $\xi = (\xi_{<40}, \xi_{40-49}, \dots, \xi_{80+})' \sim N(\mathbf{0}, \sigma_{\xi}^2 \mathbf{Q}_J^-)$  is an age-specific random effect.

Here,  $\mathbf{Q}_T$  is determined by the temporal structure and  $\mathbf{Q}_J$  is the structure matrix for the age effect. For both terms, time and age, we assume a first-order random walk prior (RW1), as we expect that the effects of contiguous age groups and the effects of contiguous time points tend to be similar. The temporal effect is supposed to be completely structured (its covariance matrix does not contain an unstructured term) because temporal trends are typically strong for most diseases (Knorr-Held, 2000).

To gain flexibility, models including different interactions are also considered. Throughout this investigation, a wide variety of models including all possible interactions were defined and fitted. For simplicity, the models proving best results using this dataset are provided. We would like to emphasize that here the goal is not to propose a model for all situations, but a battery of models based on shared components that exploit the relationship between incidence and mortality. Consequently, using other dataset a different model could be chosen. Now, we extend M1 including a gender-specific time trend (model M2). Model 3 (M3) expands M1 with an outcome-specific age term, and, finally, models 4 (M4) and 5 (M5) also broaden M1 by incorporating gender-specific temporal and age random effects, and gender-specific temporal terms and outcome-specific age effects, respectively. We would like to comment that models including outcome-specific linear trends were also studied, but they did not provide good results.

$$\text{M2} : \log r_{1igt} = \delta_g \phi_{ig} + \rho_{gt} + \xi_j, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \rho_{gt} + \xi_j, \quad (3)$$

$$\text{M3} : \log r_{1igt} = \delta_g \phi_{ig} + \gamma_t + \zeta_{1j}, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \gamma_t + \zeta_{2j}, \quad (4)$$

$$M4 : \log r_{1igt} = \delta_g \phi_{ig} + \rho_{gt} + \kappa_{gj}, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \rho_{gt} + \kappa_{gj}, \quad (5)$$

$$M5 : \log r_{1igt} = \delta_g \phi_{ig} + \rho_{gt} + \zeta_{1j}, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \rho_{gt} + \zeta_{2j}, \quad (6)$$

The vectors  $\rho$ ,  $\zeta$ , and  $\kappa$  are given next:

- $\rho = (\rho_{1,1}, \dots, \rho_{1,10}, \rho_{2,1}, \dots, \rho_{2,10})' \sim N(\mathbf{0}, \sigma_\rho^2 (\mathbf{I}_2 \otimes \mathbf{Q}_T^-))$  is a gender-specific time random effect.
- $\zeta = (\zeta_{1,<40}, \zeta_{1,40-49}, \dots, \zeta_{1,80+}, \zeta_{2,<40}, \zeta_{2,40-49}, \dots, \zeta_{2,80+})' \sim N(\mathbf{0}, \sigma_\zeta^2 (\mathbf{I}_2 \otimes \mathbf{Q}_J^-))$ ; is an outcome-specific age random effect.
- $\kappa = (\kappa_{1,<40}, \kappa_{1,40-49}, \dots, \kappa_{1,80+}, \kappa_{2,<40}, \kappa_{2,40-49}, \dots, \kappa_{2,80+})' \sim N(\mathbf{0}, \sigma_\kappa^2 (\mathbf{I}_2 \otimes \mathbf{Q}_J^-))$  is a gender-specific age random effect.

Looking at Figures 2 and 3, some kind of proportionality is observed between the distribution of crude rates by age group and the crude temporal trends. Then, it seems sensible to assume shared component models for the age and time effects. Based on this, model 6 (M6) and model 7 (M7) include shared component terms for age and time, respectively. In these cases, additional parameters  $\lambda$  and  $\zeta$  are considered for the age and time shared component terms. A detailed description of the models is provided below.

$$M6 : \log r_{1igt} = \delta_g \phi_{ig} + \rho_{gt} + \lambda \xi_j, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \rho_{gt} + \frac{1}{\lambda} \xi_j, \quad (7)$$

$$M7 : \log r_{1igt} = \delta_g \phi_{ig} + \zeta \gamma_t + \xi_j, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \frac{1}{\zeta} \gamma_t + \xi_j. \quad (8)$$

Finally, a model including spatially unstructured random effects for incidence (asymmetric formulation) is also considered. This term could explain incidence-specific variability due to region-specific factors occurring before diagnosis, such as screening, improvements in diagnostic techniques (tomography and magnetic resonance imaging in the diagnosis of brain tumors), or improvements in the completeness of the cancer registry (Ellis et al., 2014). Model 8 (M8) introduces this unstructured term as follows:

$$M8 : \log r_{1igt} = \delta_g \phi_{ig} + \rho_{gt} + \lambda \xi_j + w_{1i}, \quad \log r_{2igt} = \frac{1}{\delta_g} \phi_{ig} + \rho_{gt} + \frac{1}{\lambda} \xi_j. \quad (9)$$

In this expression,  $w_{1i}$  represents spatially unstructured random effects for incidence. Denoting by  $\mathbf{w} = (w_1, \dots, w_n)'$ , these random effects are assumed to follow a multivariate normal distribution,  $\mathbf{w} \sim N(\mathbf{0}, \sigma_w^2 \mathbf{I}_n)$ . It is noteworthy that models including other interactions (such as space-time interactions) were also considered in this work, but they did not improve results.

### 3.1 | Computation, parameter estimation, and prediction

Model fitting, inference, and prediction were carried out using Bayesian methodology, specifically, integrated nested Laplace approximations (INLA) (Rue et al., 2009). The use of this methodology is increasing in disease mapping (Blangiardo et al., 2013; Goicoa et al., 2016; Riebler et al., 2016; Schrödle & Held, 2011) due to its effectiveness and high-speed computations in latent Gaussian Markov random fields with sparse precision matrices. The computations were performed in the R version 3.6.3 (2020-03-31) (R Core Team, 2020) through the R-package *R-INLA* (Martino & Rue, 2010), version 19.05.19, on a Windows personal computer (4 × Intel Xeon Processor E5-2620 v3 (24 cores) 12 × 16GB DDR4-2133 (96GB) 3.5" SATA3 500GB). The fitting time for each model varied between 1 and 9 min approximately.

In this paper, we were interested in obtaining predictions using INLA. The reader is referred to Etxeberria et al. (2014) and Ugarte et al. (2012) to see how predicted values were obtained when the models are presented under the umbrella of generalized linear-mixed models from an empirical Bayes approach. In INLA, predictions are obtained as a part of the model-fitting itself. As prediction is the same as fitting a model with some missing data, we can simply set  $y[i] = \text{NA}$  for those unobserved values we want to predict. In our case, we were interested in getting predictions using the same

TABLE 1 DIC values for the different models

Model	M1	M2	M3	M4	M5	M6	M7	M8
	8493.778	8486.577	8401.266	8486.896	<b>8394.209</b>	8487.111	8499.073	<b>8395.496</b>

Using this dataset M5 and M8 exhibit the lowest values of DIC (in bold).

TABLE 2 Global absolute relative bias computed using one step ahead predictions

Global absolute relative bias								
	M <sub>1</sub>	M <sub>2</sub>	M <sub>3</sub>	M <sub>4</sub>	M <sub>5</sub>	M <sub>6</sub>	M <sub>7</sub>	M <sub>8</sub>
Global	0.0970	0.0910	0.0884	0.0877	0.0764	0.0877	0.1244	0.0110
Males	0.0892	0.1052	0.0701	0.0984	0.0744	0.1156	0.1218	0.0203
Females	0.1063	0.0743	0.1100	0.0750	0.0787	0.0546	0.1275	0.0481

likelihood already used to fit the data. A detailed description of how predictions were obtained caA detailed description  $n$  can be found in Appendix A.1. The full code to fit the models will be available on the GitHub of our research group (<https://github.com/spatialstatisticsupna>).

Prior distributions on the precision parameters (inverse of variance components) are required to fully specify the models. In this case, PC-priors (Simpson et al., 2017) were used for the precision parameters  $\tau_{\phi_{\text{males}}} = 1/\sigma_{\phi_{\text{males}}}^2$ ,  $\tau_{\phi_{\text{females}}} = 1/\sigma_{\phi_{\text{females}}}^2$ ,  $\tau_{\gamma} = 1/\sigma_{\gamma}^2$ ,  $\tau_{\xi} = 1/\sigma_{\xi}^2$ ,  $\tau_{\rho} = 1/\sigma_{\rho}^2$ ,  $\tau_{\zeta} = 1/\sigma_{\zeta}^2$ ,  $\tau_{\kappa} = 1/\sigma_{\kappa}^2$ , and,  $\tau_w = 1/\sigma_w^2$ . The reader is referred to Etxeberria et al. (2018) for a thorough insight into the sensitivity analysis conducted to assess the impact of different sets of hyper-priors (PC-priors, log gamma priors, and improper uniform priors on the standard deviations) on the final estimates of shared component models. In this study, sensitivity issues were not found. Besides, log gamma priors (the priors provided by default in INLA) were used for the additional parameters  $\delta_{\text{males}}$ ,  $\delta_{\text{females}}$ ,  $\lambda$ , and  $\zeta$  in the shared components. Finally, as the models do not include an intercept, sum to zero constraints were imposed in all the terms but the shared spatial effect to ensure model identifiability (Goicoa et al., 2018).

The Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) and the Watanabe–Akaike Information Criterion (WAIC) (Watanabe, 2010) were used as model selection criteria. The logarithmic score (LS) (Gneiting & Raftery, 2007) was used as an indicator of the predictive ability of the models. As suggested by one reviewer, only DIC values are displayed in Table 1, as the WAIC and LS measures ranked the models in a similar way (see Table A.1 in the Appendix). Note that using this dataset M5 and M8 exhibit the lowest values of DIC (and also WAIC and LS). To go into greater depth on the predictive performance of these models, a validation procedure is carried out in the next section.

#### 4 | VALIDATING CANCER INCIDENCE PREDICTIONS

To assess the predictive ability of all the models, one-step ahead predictions were computed based on different fitting periods. Here, as the time period used is biannual, we considered one-step ahead predictions to assess predictive ability. More precisely, the following process was used to generate predictions. Incidence predictions for the period 1997–1998 were based on models fitted in the period 1989–1996 (the minimum data we used to fit the model are four 2-year time periods). Predictions for 1999–2000 were based on data from 1989 to 1998 and so on. A total of six rounds of cross-validations were done to assess the predictive ability of the models by using the global absolute relative bias (GAR<sub>B</sub>).

$$GAR_B = \frac{\left| \sum_{igt} O_{1igt} - \sum_{igt} \hat{O}_{1igt} \right|}{\sum_{igt} O_{1igt}} \quad (10)$$

In this expression,  $O_{1igt}$  represents the observed incidence cases and  $\hat{O}_{1igt}$  is the predicted incidence cases for each area  $i$ , age group  $j$ , and time period  $t$ . To look into more detailed results, gender absolute relative biases were also computed. Results are shown in Table 2.

Figures in Table 2 clearly indicate that M8 provides the best results in terms of GAR<sub>B</sub>, with the overall bias in this model (0.011) being about 7 times lower than the second best model (M5). By gender, M8 is also the best one. For males,



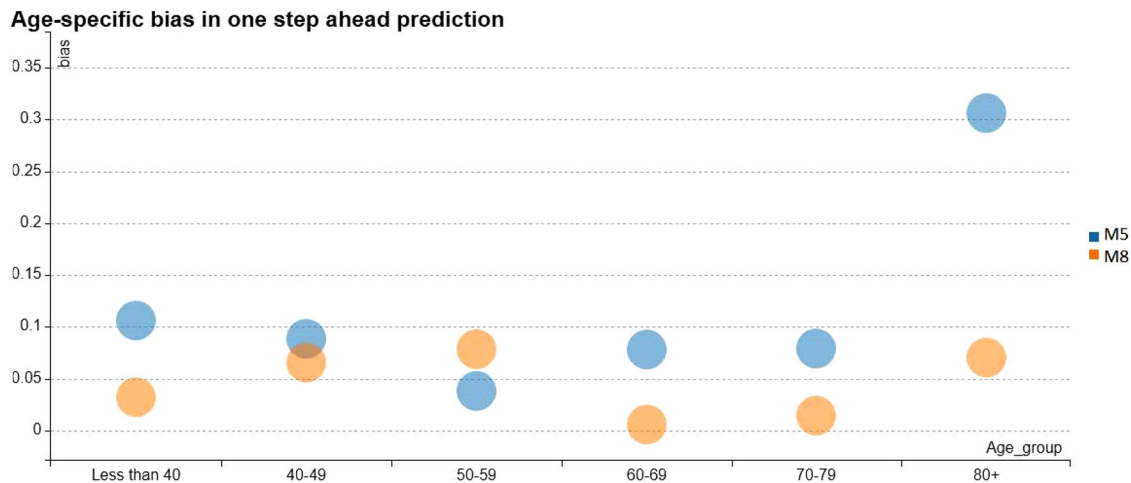


FIGURE 5 Age-specific relative biases in one-step ahead predictions

the GARB is 0.0203 (about 3 times lower than in M3 and M5, two competitive models), and for females the GARB is 0.0481, the lowest value among all models. At this point, it is important to emphasize that what differentiates M8 from the rest of the model is the age-specific shared term plus the spatially unstructured random effects for incidence. It appears that including these last terms in the model substantially improves predictive ability. More specifically, we have observed that models without spatially unstructured random effects for incidence underestimate the number of brain cancer incidence cases. Therefore, introducing this term in the model seems to improve prediction results.

Finally, as incidence varies by age group and region, it is important to assess how models predict over these groups. For health researchers, it is relevant to know if the models provide similar bias by age group and region or if there are subgroups that are better predicted than others. To gain understanding of this, age- and region-specific relative biases are computed in the next subsection.

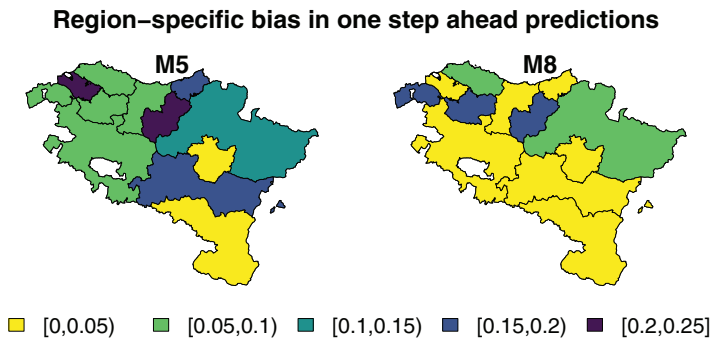
#### 4.1 | Validation by age groups and regions

Here, age-specific and region-specific relative biases are computed for each model using the following expressions:

$$\text{By age groups } \text{ARB}_j = \frac{|\sum_{igt} O_{1igt} - \sum_{igt} \hat{O}_{1igt}|}{\sum_{igt} O_{1igt}} \quad \text{By regions } \text{ARB}_i = \frac{|\sum_{gjt} O_{1igt} - \sum_{gjt} \hat{O}_{1igt}|}{\sum_{gjt} O_{1igt}}$$

Figure 5 shows interesting results on how the best two models M5 and M8 perform by age groups. In general, model M8 seems to perform best as it provides biases below 10% in all the age groups. This model provides reasonable bias results even in the more difficult age groups < 40 and 80+. For the rest of models, U-shaped biases are observed indicating a bad performance for the oldest age groups. We should not be overly concerned about providing poor predictions for the 80+ age group, as BCNS estimates in the elderly are less important than in other age groups. Brain cancer in elderly people presents some particularities. In most cases, they are not treated as they are usually asymptomatic and brain tumors in this age group have a slow growth rate. Some of the elderly patients present also multiple comorbidities, low tolerance to chemotherapy, high risk for radiation-induced neurotoxicity, and very limited life expectancies (Nayak & Iwamoto, 2010). This is the reason why brain cancer tumors are just followed up among elderly patients rather than treated. In contrast, the age group < 40 is important as brain cancer is the second most frequent cancer in children and young people after leukemia. Hence, providing good predictions is key to better organize resources for treatment and thus to avoid premature deaths (Ugarte et al., 2015b). In this age-group model, M8 performs the best.

Figure 6 gives region-specific relative biases for the best two models M5 and M8. By regions, again model M8 is clearly the best in terms of bias. Using this model, Southern Biscay and Western Gipuzkoa are the regions with the highest bias followed by Northern Biscay and Northern Navarre. In summary, model M8 would be the most suitable model for providing incidence predictions as it shows more accurate results both globally and by age groups and regions.



**FIGURE 6** Region-specific relative biases in one-step ahead predictions

**TABLE 3** Observed versus the predicted number of brain cancer incidence cases per period

Period	Males			Females				
	Observed	Fitted	95% Credible interval	Observed	Fitted	95% Credible interval		
1989–1990	196	206	182	230	159	176	155	201
1991–1992	222	223	199	253	177	182	160	209
1993–1994	274	250	220	285	189	195	174	219
1995–1996	264	261	232	302	205	218	189	252
1997–1998	254	262	235	301	218	227	198	259
1999–2000	268	264	231	302	219	240	210	272
2001–2002	234	261	230	302	237	249	216	281
2003–2004	287	286	250	324	212	238	209	271
		<b>Predicted</b>	<b>95% Credible interval</b>		<b>Predicted</b>	<b>95% Credible interval</b>		
2005–2006	-	290	249	327	-	254	218	295
2007–2008	-	302	254	346	-	255	218	296

## 5 | REAL DATA ANALYSIS

In this section, model M8 is considered to provide BCNS cancer incidence predictions in Navarre and the Basque Country by region, age group, gender, and period. This election is based on model selection criteria together with the good performance in the validation process. Using M8, we will focus on predicting incidence cases in periods when mortality figures are already available (2005–2006 and 2007–2008). First of all, the observed and the fitted number of incidence cases and their corresponding 95% credible intervals by period and gender are shown in Table 3. Predicted incidence cases (posterior means) for periods 2005–2006 and 2007–2008 in both genders and 95% credible intervals are also provided. Among males, 592 cases are predicted (290 in 2005–2006 and 302 in 2007–2008) while among females 509 are predicted (254 in 2005–2006 and 255 in 2007–2008). It can be observed that for females the fitted values are all above the observed ones. One reason may be a kind of shrinkage effect. Incidence rates for females are in general lower than in males (see Figure 3), but the difference is getting smaller with time. Hence it seems that the model tends to push female incidence towards males incidence, and hence we observed predicted incidence rates for females above the observed.

Figure 7 displays temporal incidence trends and predicted values with their 95% credible bands for 2005–2008 for both genders. This figure shows an increasing trend for both genders during the study period, and this trend could continue in the forthcoming periods for males. On the other hand, the trend seems to stabilize for women from 2005 onwards. Note that, usually the long-term forecast values present more uncertainty than forecasts for the near future. In our case, this is not relevant for two main reasons: First, under M8 both incidence and mortality have the same gender trends,  $\rho_{gt}$ , and the estimated mortality trend will be used to forecast incidence and, therefore, the uncertainty will not widen. Second, the incidence forecast is anchored around the observed mortality, which reduces uncertainty. This is a very important advantage of this modeling versus univariate incidence modeling approaches.

Figures 8 and 9 display the posterior means of predicted incidence rates for each region in the last time period (2007–2008) by age groups (rows) and gender (columns). Each region is specifically colored regarding the predicted rates per  $10^5$  inhabitants, so that it is easy to see its ranking within the different age groups and genders. To indicate the variability

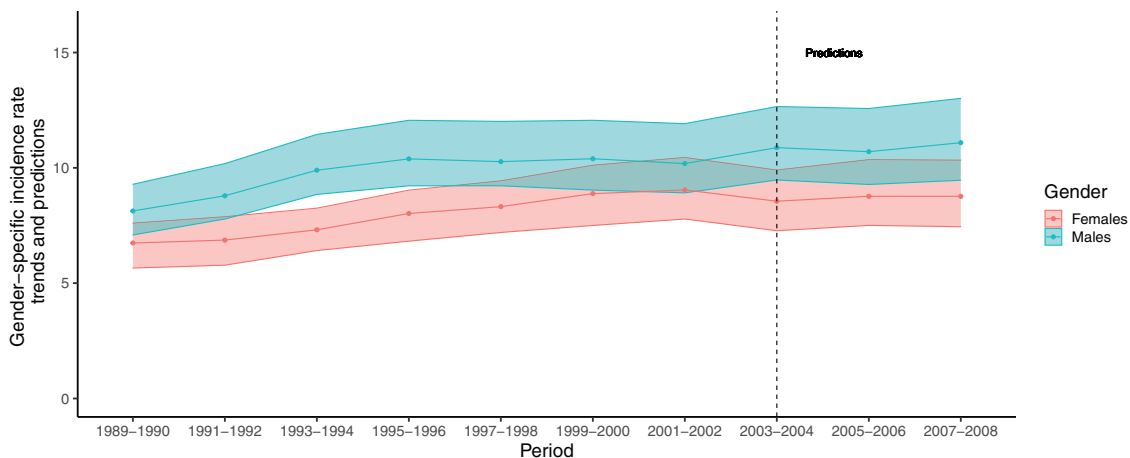


FIGURE 7 Gender-specific temporal trends and predicted rates for 2005–2006 and 2007–2008 obtained with M8

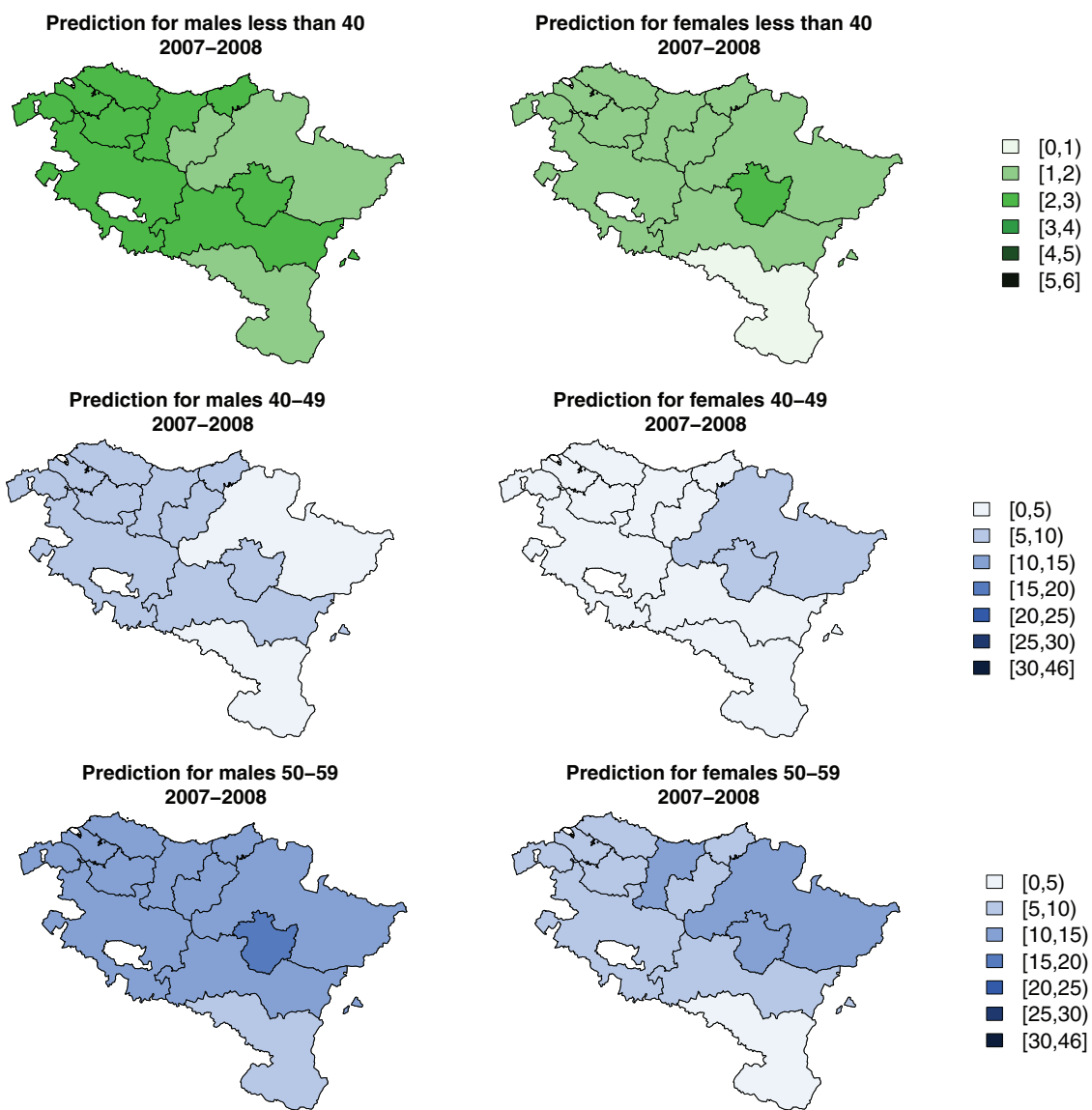


FIGURE 8 Maps of predicted incidence rates for age group < 40, 40–49, and, 50–59 for 2007–2008 period for the 11 health regions. Note that the rate scale used for < 40 is different from that used for 40–49 and, 50–59.

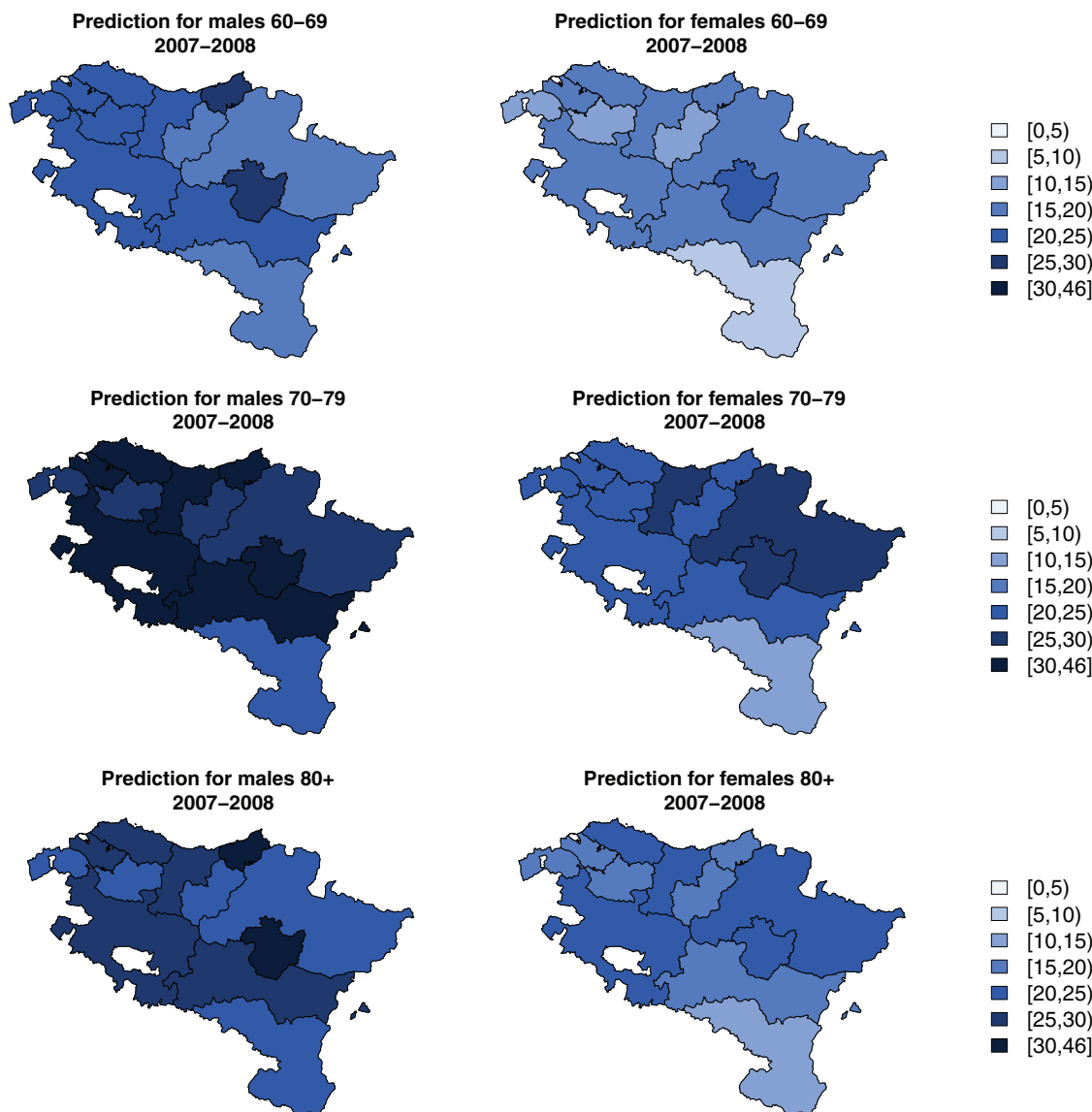


FIGURE 9 Maps of predicted incidence rates for age groups 60–69, 70–79, and 80+ for 2007–2008 period for the 11 health regions

of the predictions, maps of the coefficients of variation, are displayed in Figure A.2 in Appendix A.2. The coefficient of variation is computed as the posterior standard deviation of the predicted distribution divided by the posterior mean. One of the most important findings of this study is that neither the region nor age groups are equally affected. These maps provide valuable results as the region of Pamplona (main city of Navarre, region number 11 in Figure 1) seems to be the area with the highest rate in almost all age groups in both genders. Then, in a hypothetical brain cancer prevention plan, this area should be considered of high priority. In contrast, Southern Navarre (region number 9 in Figure 1) is the region with the lowest rate for most age groups and both genders.

Little variation within regions is observed for age groups < 40, 40–49, and 50–59 where rates remain below 20 cases per  $10^5$  (below 5 cases per  $10^5$  for < 40). For the 60–69 age group, males living in Pamplona and Donostia-Bajo Bidasoa, the capital city of the province of Gipuzkoa, are the most affected.

Brain cancer rates reach their maximum in the 70–79 age group, in which some geographical differences are found. Regions located on the coast of the Bay of Biscay, Alava, Pamplona, and Mid-Navarre are the ones with the highest rates in males. In females, West Gipuzkoa, Navarra North, and Pamplona are the areas with the highest rates. In both genders, Southern Navarre is the one with the lowest rates. Finally, rates decrease slightly for the 80+ age group with maps more similar to those for the 60–69 age group.

## 6 | DISCUSSION

High-quality and preferably long-term population-based data on cancer incidence and mortality are crucial for cancer control and prevention. Compared with mortality figures, incidence cases are usually available after approximately 3 years due to administrative and procedural delays. Consequently, health policymakers consider alternative information, usually relying on predictions based on statistical models. Approaches based on age-period-cohort models are usually employed in the literature to provide predictions of cancer mortality or incidence counts, but these methods are not useful for rare and lethal cancers such as BCNS or pancreatic cancer due to data scarcity. Our proposal comes to fill this gap. In this paper, gender- and age-specific shared component models are proposed to predict incidence when mortality is already available. The high correlation between incidence and mortality in brain cancer supports the joint modeling of both processes increasing the effective sample size. The major advantage of our method is that it elegantly exploits the correlation between incidence and mortality allowing disaggregated predictions by region, age groups, and gender, variables playing an important role in BCNS epidemiology (Miranda-Filho et al., 2016). This would be impossible if a univariate prediction model for incidence had been considered due to the scarce number of cases in certain regions and age groups.

Although model-based predictions should be interpreted in light of data limitations and modeling assumptions, we found that our proposed model provides accurate results (with a posterior coefficient of variations under 15%) in general, and in particular in the sensitive age group < 40. Brain tumors are an important type of cancer in children and young adults, and understanding their epidemiology is essential for clinicians and for those involved in the care of patients or investigating the cause of primary brain tumors in these age groups (McNeill, 2016). It should be noted that only a small proportion of brain tumors can be explained by established risk factors (exposure to ionizing radiation, rare mutations of penetrant genes, and familial history) (Fisher et al., 2007).

We expect that predictions at a very disaggregated level will contribute to complete the cancer data series improving health system planning and management of lethal cancers. The results presented in this study also indicate important regional variations in BCNS incidence predictions among Navarre and Basque Country. Projected gender-specific trends indicate that males will have higher incidence rates of BCNS than females. This is consistent with the results obtained in other regions in which the male-to-female ratio ranges from 1.0 to 2.7 (Miranda-Filho et al., 2016). It has been suggested that gender differences could be due to sex hormones and genetic features (McKinley et al., 2000). Like any forecasting method, our proposal also has some limitations. First, not all the regions and age groups are predicted equally well. Data scarcity in some age groups is really an obstacle to provide accurate predictions. Second, the predicted trends are based on the observed ones that do not capture the effects of future events. For example, the implementation of new screening programs, improvements in the data registration, or any change in the definition of a particular malignancy could affect the number of incidence cases to a large extent. Finally, we are aware that our data are not very updated but unfortunately we do not have access to more recent incidence data yet. However, despite these limitations, the methodology presented in this article is a promising alternative to existing techniques when predicting rare and lethal cancer types by age, gender, and region. This paper will provide regional cancer registries with a valuable predictive tool.

### ACKNOWLEDGMENTS

We would like to thank Eva Ardanaz from the Navarre Cancer Registry (Public Health and Labor Institute of Navarre), Nerea Larrañaga from the Basque Cancer Registry, and Covadonga Audicana from the Basque Mortality Registry for providing the data. The work has been supported by Project PID2020-113125RB-I00, MCIN/AEI /10.13039/501100011033, and European Union NextGenerationEU/PRTR and Proyecto Jóvenes Investigadores PJUPNA2018-11.


### CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

### DATA AVAILABILITY STATEMENT

Synthetic data comparable to the original data in size and structure have been included.

### 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

Maria D. Ugarte  <https://orcid.org/0000-0002-3505-8400>

## REFERENCES

- Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20.
- Bezerra-de Souza, D. L., Bernal, M. M., Gómez, F. J., & Gómez, G. J. (2012). Predictions and estimations of colorectal cancer mortality, prevalence and incidence in Aragon, Spain, for the period 1998–2022. *Revista Española de Enfermedades Digestivas*, 104(10), 518–523.
- Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 4, 33–49.
- Bray, F., Kohler, B., & Ferlay, J. (2013). Data comparability and quality. *Cancer Incidence in Five Continents 2013*; 10: 89–106.
- Connelly, J. M., & Malkin, M. G. (2007). Environmental risk factors for brain tumors. *Current Neurology and Neuroscience Reports*, 7(3), 208–214.
- Dyba, T., & Hakulinen, T. (2000). Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Statistics in Medicine*, 19(13), 1741–1752.
- Ellis, L., Woods, L. M., Estève, J., Eloranta, S., Coleman, M. P., & Rachet, B. (2014). Cancer incidence, survival and mortality: Explaining the concepts. *International Journal of Cancer*, 135(8), 1774–1782.
- Etxeberria, J., Goicoa, T., López-Abente, G., Riebler, A., & Ugarte, M. D. (2017). Spatial gender-age-period-cohort analysis of pancreatic cancer mortality in Spain (1990–2013). *PloS ONE*, 12(2), e0169751.
- Etxeberria, J., Goicoa, T., & Ugarte, M. (2018). Joint modelling of brain cancer incidence and mortality using Bayesian age-and gender-specific shared component models. *Stochastic Environmental Research and Risk Assessment*, 32(10), 2951–2969.
- Etxeberria, J., Goicoa, T., Ugarte, M. D., & Militino, A. F. (2014). Evaluating space-time models for short-term cancer mortality risk predictions in small areas. *Biometrical Journal*, 56(3), 383–402.
- Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., Gavin, A., Visser, O., & Bray, F. (2018). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*, 103, 356–387.
- Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., & Bray, F. (2013). Cancer incidence and mortality worldwide. IARC CancerBase No. 11. IARC.
- Fisher, J. L., Schwartzbaum, J. A., Wrensch, M., & Wiemels, J. L. (2007). Epidemiology of brain tumors. *Neurologic Clinics*, 25(4), 867–890.
- Forman, D., Bray, F., Brewster, D., Gombe Mbalawa, C., Kohler, B., Piñeros, M., Steliarova-Foucher, E., Swaminathan, R., & Ferlay, J. (2013). *Cancer incidence in five continents*, vol. X (electronic version). Lyon: IARC.
- Galceran, J., Ameijide, A., Carulla, M., Mateos, A., Quirós, J. R., Rojas, D., Alemán, A., Torrella, A., Chico, M., Vicente, M., Díaz, J. M., Larrañaga, N., Marcos-Gragera, R., Sánchez, M. J., Perucha, J., Franch, P., Navarro, C., Ardanaz, E., Bigorra, J., ... REDECAN Working Group. (2017). Cancer incidence in Spain, 2015. *Clinical and Translational Oncology*, 19(7), 799–825.
- Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.
- Goicoa, T., Adin, A., Ugarte, M., & Hodges, J. S. (2018). In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment*, 32, 749–770. <https://doi.org/10.1007/s00477-017-1405-0>
- Goicoa, T., Ugarte, M., Etxeberria, J., & Militino, A. F. (2016). Age–space–time car models in Bayesian disease mapping. *Statistics in Medicine*, 35(14), 2391–2405.
- Hakulinen, T., Teppo, L., & Saxén, E. (1986). Do the predictions for cancer incidence come true? Experience from Finland. *Cancer*, 57(12), 2454–2458.
- Holford, T. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, 311–324.
- Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18), 2555–2567.
- Knorr-Held, L., & Best, N. G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(1), 73–85.
- MacNab, Y. C. (2010). On Bayesian shared component disease mapping and ecological regression with errors in covariates. *Statistics in Medicine*, 29(11), 1239–1249.
- MacNab, Y. C. (2016a). Linear models of coregionalization for multivariate lattice data: A general framework for coregionalized multivariate car models. *Statistics in Medicine*, 35(21), 3827–3850.
- MacNab, Y. C. (2016b). Linear models of coregionalization for multivariate lattice data: Order-dependent and order-free cmcars. *Statistical Methods in Medical Research*, 25(4), 1118–1144.
- Martino, S., & Rue, H. (2010). *Implementing approximate Bayesian inference using integrated nested Laplace approximation: A manual for the INLA program*. Department of Mathematical Sciences, NTNU, Norway.
- McKinley, B. P., Michalek, A. M., Fenstermaker, R. A., & Plunkett, R. J. (2000). The impact of age and gender on the incidence of glial tumors in New York State from 1976–1995. *Journal of Neurosurgery*, 93(6), 932–939.
- McNeill, K. A. (2016). Epidemiology of brain tumors. *Neurologic Clinics*, 34(4), 981–998.
- Miranda-Filho, A., Piñeros, M., Soerjomataram, I., Deltour, I., & Bray, F. (2016). Cancers of the brain and CNS: Global patterns and trends in incidence. *Neuro-Oncology*, 19(2), 270–280. <https://doi.org/10.1093/neuonc/now166>
- Møller, B., Fekjær, 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. *Statistics in Medicine*, 22(17), 2751–2766.

- Nayak, L., & Iwamoto, F. M. (2010). Primary brain tumors in the elderly. *Current Neurology and Neuroscience Reports*, 10(4), 252–258.
- Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology*, 14(1), 124–129.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. <https://www.R-project.org/>
- Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4), 1145–1165.
- 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.
- Sánchez, M., Payer, T., De Angelis, R., Larrañaga, N., Capocaccia, R., Martínez, C., & CIBERESP Working Group. (2010). Cancer incidence and mortality in Spain: Estimates and projections for the period 1981–2012. *Annals of Oncology*, 21(Suppl\_3), iii30–iii36.
- Schrödle, B., & Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics*, 22(6), 725–734.
- Simpson, D., Rue, H., Riebler, A., & Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1), 1–28.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.
- Steliarova-Foucher, E., Colombet, M., Ries, L. A., Moreno, F., Dolya, A., Bray, F., Hesselting, P., Shin, H. Y., Stiller, C. A., & IICC-3 contributors. (2017). International incidence of childhood cancer, 2001–10: A population-based registry study. *The Lancet Oncology*, 18(6), 719–731.
- Teppo, L., Hakulinen, T., & Saxen, E. (1974). The prediction of cancer incidence in Finland for the year 1980 by means of cancer registry material. *Annals of Clinical Research*, 1974, 112–5.
- Ugarte, M., Adin, A., Goicoa, T., Casado, I., Ardanaz, E., & Larrañaga, N. (2015a). Temporal evolution of brain cancer incidence in the municipalities of Navarre and the Basque Country, Spain. *BMC Public Health*, 15(1), 1018.
- Ugarte, M., Adin, A., Goicoa, T., & López-Abente, G. (2015b). Analyzing the evolution of young people's brain cancer mortality in Spanish provinces. *Cancer Epidemiology*, 39(3), 480–485.
- Ugarte, M. D., Goicoa, T., Etxeberria, J., & Militino, A. F. (2012). Projections of cancer mortality risks using spatio-temporal P-spline models. *Statistical Methods in Medical Research*, 21(5), 545–560.
- Uhry, Z., Colonna, M., Remontet, L., Grosclaude, P., Carr, N., Couris, C. M., & Velten, M. (2007). Estimating infra-national and national thyroid cancer incidence in France from cancer registries data and national hospital discharge database. *European Journal of Epidemiology*, 22(9), 607–614.
- Watanabe S., (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec), 3571–3594.

## 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:** Etxeberria, J., Goicoa, T., & Ugarte, M. D. (2022). Using mortality to predict incidence for rare and lethal cancers in very small areas. *Biometrical Journal*, 00–00. <https://doi.org/10.1002/bimj.202200017>

## APPENDIX

### A.1 | Model Parameter estimation and prediction using INLA

Let us define  $\theta = (\phi^*, \gamma, \delta, \dots)$  the vector of parameters assuming a Gaussian Markov random field prior to  $\theta$  with mean  $\mathbf{0}$  and precision matrix  $\mathbf{G}$  which depends on some hyperparameters  $\lambda_k$ .

The objectives of the Bayesian computation are the following posterior marginal distributions (p.m.d):

$$\begin{aligned}
 \text{A } \pi(\theta_i | \mathbf{y}) &= \int_{\lambda} \underbrace{\pi(\theta_i | \lambda, \mathbf{y})}_{(A)} \underbrace{\pi(\lambda | \mathbf{y})}_{(B)} d\lambda \quad \text{p.m.d. of the parameters vector } \theta. \\
 \text{B } \pi(\lambda_k | \mathbf{y}) &= \int_{\lambda} \underbrace{\pi(\lambda | \mathbf{y})}_{(B)} d\lambda_{-k} \quad \text{p.m.d of the hyperparameter } \lambda_k.
 \end{aligned}$$

Thus, we need to compute (A) and (B)

**Term (B)** The approximation of the posterior marginal distribution of the hyper-parameters  $\lambda$  is given by

$$\tilde{\pi}(\lambda | \mathbf{y}) \propto \frac{\pi(\theta, \lambda, \mathbf{y})}{\tilde{\pi}(\theta | \lambda, \mathbf{y})} \Big|_{\theta = \theta^*(\lambda)}, \quad (\text{A.1})$$

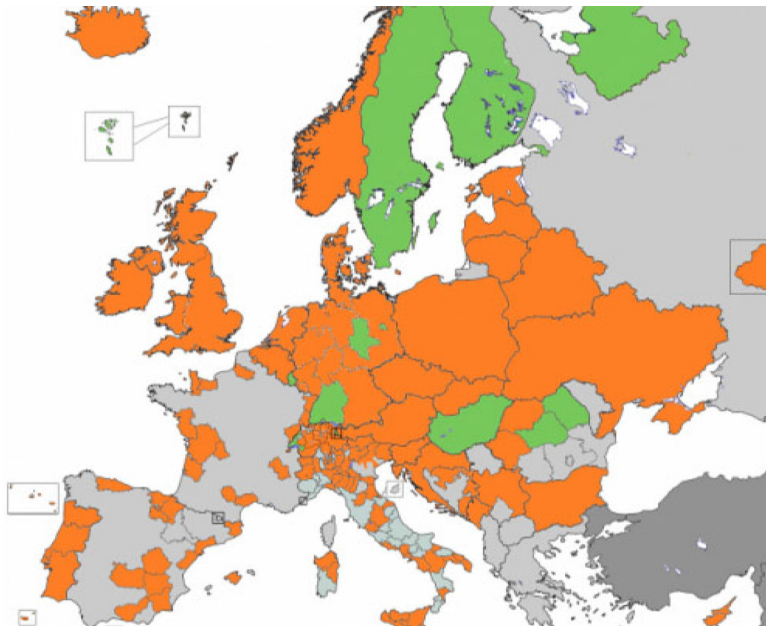


FIGURE A.1 Geographic distribution of the general cancer registries (all cancer sites and all ages) available in Europe. Figure available at The European Network of Cancer Registries-Joint Research Centre project. (See <https://www.enrcr.eu/> for more detail. Last accessed November 2021).

TABLE A.1 Model selection criteria and predictive ability for the different models.

Model	DIC	DIC incidence	DIC mortality	Dbar	pD	pD incidence	pD mortality	WAIC	LS
M1	8493.778	4010.64	4483.13	8461.169	32.608	14.11	18.50	8502.196	1.789
M2	8486.577	4007.49	4479.09	8449.097	37.480	16.06	21.42	8496.557	1.788
M3	8401.266	3976.18	4425.08	8363.698	37.568	16.97	20.59	8409.711	1.770
M4	8486.896	4006.26	4480.64	8444.421	42.474	18.25	24.23	8498.831	1.788
M5	8394.209	3973.10	4421.11	8351.786	42.423	18.95	23.47	8403.822	1.768
M6	8487.111	4004.72	4482.39	8452.065	35.046	15.00	20.04	8496.891	1.788
M7	8499.073	4007.90	4491.18	8465.631	33.441	14.32	19.12	8508.602	1.791
M8	8395.496	3970.43	4425.06	8351.438	44.058	20.78	23.27	8404.358	1.769

Abbreviations: DIC, Deviance Information Criterion; LS, logarithmic score; WAIC, Watanabe–Akaike Information Criterion; Dbar, Posterior mean of the deviance; pD, Effective number of parameters in the model.

where  $\tilde{\pi}(\theta|\lambda, \mathbf{y})$  is the Gaussian approximation of the full conditional distribution of  $\theta$  (Rue & Held, 2005) and  $\theta^*(\lambda)$  is the mode of the full conditional distribution of  $\theta$  for a given  $\lambda$ .

**Term (A)** Rue et al. (2009) propose three approaches to approximate  $\pi(\theta_i|\lambda, \mathbf{y})$ : (1) a Gaussian approximation, (2) a full Laplace approximation, and (3) a simplified Laplace approximation. According to Rue and Martino (2007), the Gaussian approximation gives quite satisfactory results in a short time.

Once both terms are approximated, the p.m.d. of the **parameters vector**  $\theta$  is given by

$$\tilde{\pi}(\theta|\mathbf{y}) = \sum_k \tilde{\pi}(\theta|\lambda_k, \mathbf{y}) \tilde{\pi}(\lambda_k|\mathbf{y}) \Delta_k, \quad (\text{A.2})$$

where  $\Delta_k$  is an area weight assigned to each  $\lambda_k$ . Its size depends on the actual strategy of choosing appropriate  $\lambda_k$ s. The INLA approach is available in an R package named R-INLA ([www.r-inla.org](http://www.r-inla.org)).

To compute the prediction, we rely on the posterior predictive distribution given by

$$\tilde{\pi}(y_{\text{miss}}|\mathbf{y}) \propto \tilde{\pi}(y_{\text{miss}}|\theta) \tilde{\pi}(\theta|\mathbf{y}) \quad (\text{A.3})$$

- $\tilde{\pi}(y_{\text{miss}}|\theta)$  is the likelihood function (the same as for the observed  $\mathbf{y}$ ).
- $\tilde{\pi}(\theta|\mathbf{y})$  is the posterior distribution of the parameter vector  $\theta$  given  $\mathbf{y}$  computed by (A).



## A.2 | Coefficients of variation of the predicted rates by regions and age groups obtained as the posterior standard deviation of the rates divided by the posterior mean

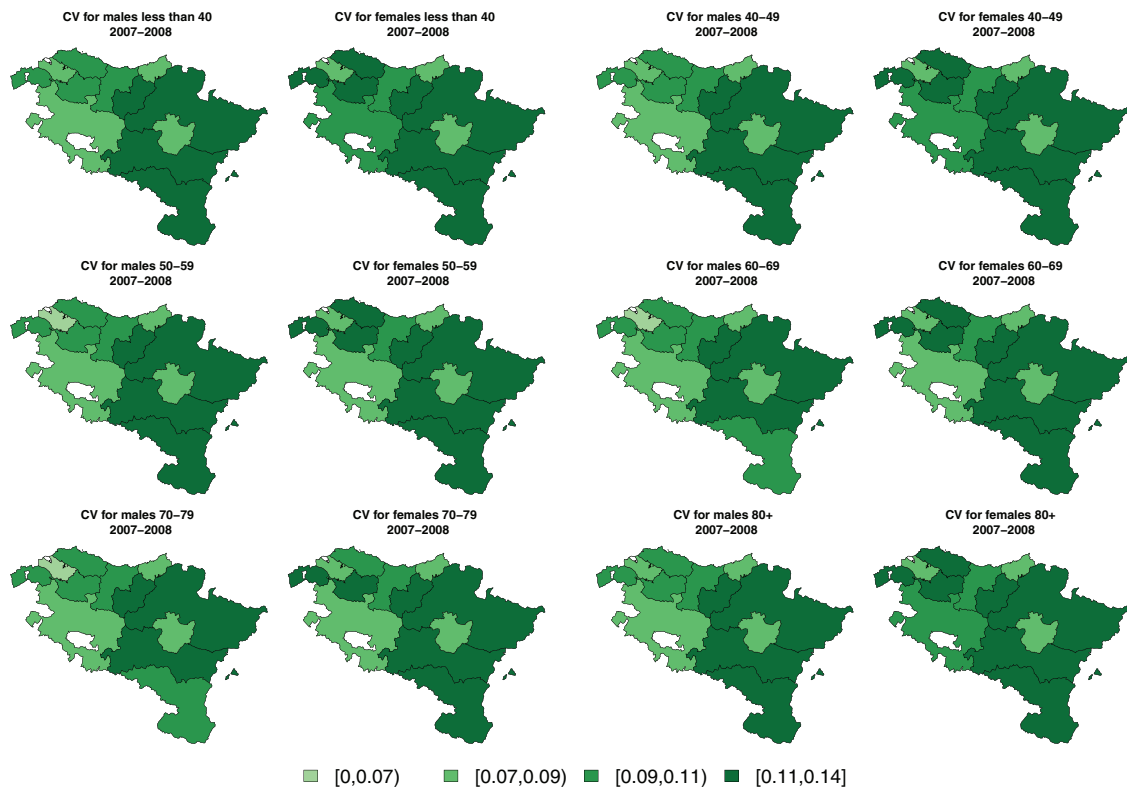


FIGURE A.2 Coefficients of variation by regions and age groups obtained as the posterior standard deviation of the rates divided by the posterior mean.