




## Methods

# Multidimensional penalized splines for incidence and mortality-trend analyses and validation of national cancer-incidence estimates

Zoé Uhry <sup>1,2\*</sup>, Edouard Chatignoux,<sup>1</sup> Emmanuelle Dantony,<sup>2,3</sup> Marc Colonna,<sup>4</sup> Laurent Roche,<sup>2,3</sup> Mathieu Fauvernier,<sup>2,3</sup> Gautier Defosse,<sup>5</sup> Sandra Leguyader-Peyrou,<sup>6</sup> Alain Monnereau,<sup>6</sup> Pascale Grosclaude,<sup>7,8</sup> Nadine Bossard,<sup>2,3</sup> Laurent Remontet<sup>2,3</sup> and the French network of cancer registries (Francim)

<sup>1</sup>Direction des Maladies Non Transmissibles et des Traumatismes, Santé Publique France, Saint-Maurice, France, <sup>2</sup>Service de Biostatistique-Bioinformatique, Pôle Santé Publique, Hospices Civils de Lyon, Lyon, France, <sup>3</sup>Laboratoire de Biométrie et Biologie Évolutive, UMR 5558, CNRS, Université Lyon 1, Université de Lyon, Villeurbanne, France, <sup>4</sup>Registre des cancers de l'Isère, Grenoble, France, <sup>5</sup>Registre des cancers du Poitou-Charentes, Poitiers, France, <sup>6</sup>Registre des hémopathies malignes de la Gironde, Institut Bergonié, Bordeaux, France, <sup>7</sup>Registre des cancers du Tarn Cancer, Institut Claudius Regaud, Institut universitaire du cancer de Toulouse Oncopole (IUCT-O), Toulouse, France and <sup>8</sup>Laboratoire d'Epidémiologie et Analyses en Santé Publique (LEASP), UMR 1027, Inserm; Université Toulouse III, Toulouse, France

\*Corresponding author. Service de Biostatistique-Bioinformatique Centre Hospitalier Lyon Sud 165 Chemin du Grand Revoyet 69495 Pierre-Bénite Cedex, France. E-mail: zoe.uhry@chu-lyon.fr

Editorial decision 7 April 2020; Accepted 14 April 2020

## Abstract

**Background:** Cancer-incidence and mortality-trend analyses require appropriate statistical modelling. In countries without a nationwide cancer registry, an additional issue is estimating national incidence from local-registry data. The objectives of this study were to (i) promote the use of multidimensional penalized splines (MPS) for trend analyses; (ii) estimate the national cancer-incidence trends, using MPS, from only local-registry data; and (iii) propose a validation process of these estimates.

**Methods:** We used an MPS model of age and year for trend analyses in France over 1990–2015 with a projection up to 2018. Validation was performed for 22 cancer sites and relied essentially on comparison with reference estimates that used the incidence/health-care ratio over the period 2011–2015. Alternative estimates that used the incidence/mortality ratio were also used to validate the trends.

**Results:** In the validation assessment, the relative differences of the incidence estimates (2011–2015) with the reference estimates were <5% except for testis cancer in men and < 7% except for larynx cancer in women. Trends could be correctly derived since 1990 despite incomplete histories in some registries. The proposed method was applied to estimate

the incidence and mortality trends of female lung cancer and prostate cancer in France.

**Conclusions:** The validation process confirmed the validity of the national French estimates; it may be applied in other countries to help in choosing the most appropriate national estimation method according to country-specific contexts. MPS form a powerful statistical tool for trend analyses; they allow trends to vary smoothly with age and are suitable for modelling simple as well as complex trends thanks to penalization. Detailed trend analyses of lung and prostate cancers illustrated the suitability of MPS and the epidemiological interest of such analyses.

**Key words:** incidence, mortality, penalized splines, generalized additive models, trend analyses, cancer, cancer registry

### Key Messages

- Detailed cancer-incidence and mortality-trend analyses are essential to gain epidemiological insights but require appropriate statistical modelling.
- Multidimensional penalized splines (MPS) form a powerful statistical tool for incidence and mortality-trend analyses. MPS allow the trends to vary smoothly with age and may model simple as well as complex trends through penalization, which provides the 'best' trade-off between fit and smoothness.
- In countries without a nationwide cancer registry, an additional issue is to produce and validate national estimates from local data.
- In the French context, it is possible to obtain valid national cancer-incidence trends directly from local-registry data that cover 20% of the population, without any correction and including registries with incomplete histories.
- The validation process is based on comparison with reference estimates that uses health-care data for recent estimates and with alternative estimates that use mortality for historical trend. This validation process may be applied in other countries to help in choosing the most appropriate national estimation method according to each country's specific context.

## Introduction

Detailed incidence and mortality-trend analyses are key elements in epidemiological surveillance programmes, especially in the cancer field, which benefits from many population-based registry incidence data.<sup>1</sup> Trends according to year of diagnosis or birth cohort detailed by age may be linked to the dynamics of known risk factors or raise hypotheses about emerging unconfirmed factors.

For such refined trend analyses, appropriate modelling is required and the general additive model (*gam*) framework as developed by Wood<sup>2,3</sup> is an appealing opportunity, especially multidimensional penalized-regression splines (MPS). MPS are flexible models that provide smooth rates and allow the trends to vary with age; they are suitable for modelling simple as well as complex trends. MPS are implemented in the R package *mgcv* and are now a mature tool, both theoretically and numerically.<sup>3,4</sup> MPS thus represent a promising modelling perspective for cancer-incidence and mortality-trend analyses, which are still rarely used.

Besides this modelling aspect, when the objective is to study national trends, the many countries in which cancer registries cover only a part of the population face the issue of estimating national incidence from local data.<sup>1</sup> National cancer incidence ( $I_N$ ) is then usually estimated by correcting registry-area incidence ( $I_R$ ) using the ratio between national and registry-area mortality:  $I_N = I_R * M_N / M_R = M_N * I_R / M_R$ . Different versions of such I/M approaches were proposed.<sup>1,5-8</sup> An additional common issue is that several registries may not have historical data and, in this case, only recent estimates are usually derived.<sup>1</sup> Few methods have attempted to estimate national trends from local-registry data in such cases; one may cite the specific I/M approaches developed in France<sup>6,7,9</sup> or in Spain.<sup>5</sup>

Although I/M approaches proved very useful<sup>1,5-8</sup> and valid,<sup>10</sup> they have some limitations. First, with histological codes not being available in mortality data, I/M approaches are not applicable for various hematologic malignancies or cancer histological subtypes. Second, mortality became less informative about the incidence for a few cancers (thyroid, melanoma, prostate, etc.). Third, the

$M_N/M_R$  correcting factor inflates the variability of the  $I_N$  estimate, especially when mortality is low, and, fourth, confidence intervals may be difficult to derive. In France, as coverage of the registry area is increasing, these limitations led us to question the use of an  $I/M$  approach to estimate the national cancer incidence.

For all these reasons, we adopted a new method to estimate the French national cancer-incidence and mortality trends over 1990–2018. The present work therefore had three objectives:

- i. to promote and illustrate the use of MPS for incidence and mortality-trend analyses,
- ii. to propose a method to estimate the national cancer-incidence trend, using MPS, from only local-registry data,
- iii. to propose a validation process for these national incidence estimates, based on comparison with alternative estimates obtained using health-care or mortality data.

The paper focuses on the examples of lung cancer in women and prostate cancer; however, validation results are provided for 22 cancer sites in the [Supplementary data](#), available at *IJE* online.

## Methods

### Incidence, mortality and population data

Incidence data from 1975 to 2015 were provided by the French cancer registries, which cover 19 to 22 districts (*Départements*), depending on the cancer site, i.e. 21–24% of the French population. The geographical area covered by these districts will be referred to herein as the ‘registry area’. Note that several registries do not have historical data (first available year ranging from 1975 to 2009; [Supplementary Figure S1](#), available as [Supplementary data](#) at *IJE* online) and thus cancer incidence in this registry area is fully observed only since 2009. French national mortality from 1975 to 2015 was provided by the *Centre d'épidémiologie sur les causes médicales de décès (CépiDc-Inserm)*. The person-years were calculated by sex, district, annual age and year (1975–2018) from official population data. To analyse trends over 1990–2018, data from 1985 were used for incidence (to stabilize estimation in 1990) and from 1975 for mortality (to estimate long-term cohort indicators not presented here). Note that 2018 estimates thus result from a short projection. For prostate-cancer incidence, though, no projection was performed due to the high uncertainty about its short-term evolution (the last year shown is 2015). In this paper, the year will refer to the year of diagnosis for incidence (respectively death for mortality) and the cohort will refer to the year of birth.

### Health-care data for external validation

For each cancer site and for the period 2011–2015, three indicators were derived from the hospitalization data and health-insurance data (agreements for full reimbursement of medical costs)<sup>11</sup>: (i) number of newly hospitalized patients (i.e. not hospitalized in the 2 previous years,  $H$ ); (ii) number of patients who obtained a first agreement from their health insurance ( $A$ ); and (iii) number of patients newly hospitalized or who obtained a first agreement ( $HA$ ).

### Cancer sites studied

Twenty-two cancer sites were studied ([Supplementary Table 1](#), available as [Supplementary data](#) at *IJE* online) and all analyses were performed separately by sex and site. The paper focus on the examples of female lung cancer and prostate cancer; however, validation results are provided for all cancer sites in the [Supplementary data](#), available at *IJE* online.

### Introduction to MPS and the model for national mortality-trend analyses, 1990–2018

National mortality rates by year and annual age were modelled in a Poisson regression by an MPS of age and year, called *tensor* and denoted by  $te(a, y)$ <sup>2,3</sup>:

$$D_{a,y} \sim \text{Poisson}(\mu_{a,y} \cdot PY_{a,y}) \text{ and } \text{Log}(\mu_{a,y}) = te(a, y),$$

Model 1

where  $D_{a,y}$  is the number of cancer deaths in France (age  $a$ , year  $y$ ),  $\mu_{a,y}$  is the mortality rate and  $PY_{a,y}$  is the person-years.  $te(a, y)$  is derived from two functions, namely  $f(a)$  and  $g(y)$  (e.g. splines) called the *marginal basis*, which represents the effect of age and year, respectively;  $te(a, y)$  is obtained by multiplying term by term these marginal bases (tensor product) and has  $M \times L$  parameters if the bases have  $M$  and  $L$  parameters, respectively (see Appendix A1 for details).

MPS allow modelling the potentially complex effects of age and year (non-linear effects and interactions). This model has many parameters and thus, to avoid overfitting, these parameters are estimated by maximizing a penalized likelihood, which makes a trade-off between the fit and the smoothness of the rates obtained. This trade-off is controlled by smoothing the parameters that are estimated automatically. In Model 1, there are two smoothing parameters, one for each direction of age and year, so that the predicted rate at a fixed age varies smoothly with year and reciprocally.<sup>3</sup> If the penalization is strong, then corresponding effects will be linear.

An usual choice for marginal bases is restricted cubic splines.<sup>12</sup> One general principle when using MPS is to

choose a number of knots of these splines slightly higher than deemed necessary and let the penalization avoid overfitting.<sup>3</sup> Knots were placed every 5 years for  $g(y)$  and every 10 years for  $f(a)$  (see [Supplementary Table 2](#), available as [Supplementary data](#) at *IJE* online).

### Model to estimate national incidence trends, 1990–2018

To estimate the national incidence in France, we fitted the following model using data from all registries:

$$K_{j,a,y} \sim \text{Poisson}(\lambda_{j,a,y} \cdot PY_{j,a,y}) \text{ and} \\ \text{Log}(\lambda_{j,a,y}) = \text{te}_\lambda(a, y) + u_j, \text{ with } u_j \sim N(0, \sigma^2). \quad \text{Model 2}$$

In this model,  $K_{j,a,y}$  is the number of cancer cases (district  $j$ , age  $a$ , year  $y$ ),  $\lambda_{j,a,y}$  is the incidence rate,  $PY_{j,a,y}$  is the corresponding person-years and  $u_j$  is the district random effect.  $\sigma$  represents the incidence variability between the districts. The knots were chosen as in Model 1, except for breast, prostate and ‘all cancers’, which required a higher flexibility (see [Supplementary Table 2](#), available as [Supplementary data](#) at *IJE* online). R codes for Model 1 and Model 2 are available at [https://github.com/uhryzoe/MPS\\_IncidenceTrends](https://github.com/uhryzoe/MPS_IncidenceTrends).

With a district random effect, Model 2 is naturally designed to infer national incidence from a sample of districts, including correct variance estimation (the lower the incidence varies between districts, the better the precision of the national estimates). National incidence is estimated as the marginal incidence, i.e. the average incidence over the distribution of the district random effects:

$$\hat{\lambda}_{a,y}^N = \exp(\hat{\text{te}}_\lambda(a, y)) \cdot \exp\left(\frac{\hat{\sigma}^2}{2}\right). \quad \text{Formula 1}$$

This method of estimating national incidence relies on three fundamental assumptions:

- i. National incidence may be currently estimated from registry incidence data; this requires that districts with a registry be a random sample of French districts and, thus, that incidence in the registry area be close to the whole-of-France incidence ( $I_R \sim I_N$ ).
- ii. This equality  $I_R \sim I_N$  holds over the whole study period; i.e. the incidence trend in the registry area is identical to the incidence trend in the whole of France (trend  $I_R \sim \text{trend } I_N$ ).
- iii. The incidence trend in the registry area is correctly estimated despite incomplete histories in some registries.

### Validation of the method

**Validation of the first assumption: national incidence may currently be estimated from the incidence data of registries**  
In France, there is no gold standard for national incidence estimation. However, accurate national estimates may be obtained using health-care (HC) data, which are more valuable proxies of incidence<sup>11,13</sup> than mortality.<sup>14</sup> We derived such national estimates within the context of district-level cancer-incidence prediction in France, using a calibration model that proved to provide unbiased estimates.<sup>11,15</sup> This model is detailed in Appendix A2 (Model 3); briefly, the HCl ratio is first modelled according to age in the registry area and this age-specific ratio is then applied to the age-specific number of the national HC data to derive national incidence. However, due to the lack of history of HC data, this approach could not be used over the whole study period and we thus focused on a recent period, namely 2011–2015. Three distinct national estimations, considered here as reference estimates, were carried out with the calibration model using the HA/I, H/I or A/I ratio, respectively. The comparison of our national estimates (Formula 1) with these references is the key validation element.

Furthermore, for a correct variance estimation of the national incidence obtained from Model 2, the districts with a registry have to be a random sample from all French districts in terms of incidence rates (i.e. they have to spread across the range of all possible values). This assumption may be indirectly examined by looking at the distribution over French districts of the HC rates.

### Validation of the second assumption: trend $I_R \sim \text{trend } I_N$

The incidence trend being unobservable (even in the registry area, due to incomplete histories), this second assumption was indirectly assessed by comparing graphically the mortality trends in France and in the registry area from 1990 to 2015.

### Validation of the third assumption: trend $I_R$ correctly estimated despite incomplete histories

Because several registries have incomplete histories (see [Supplementary Figure 1](#), available as [Supplementary data](#) at *IJE* online), there is no straightforward way to verify that the incidence trend in the registry area is well estimated from Model 2. Nevertheless, as a sensitivity analysis, we performed an alternative estimate of the trend  $I_R$  using a model that draws information from mortality to predict the incidence in each district (Model 4; see Appendix A3), thus providing more reliable historical estimates for  $I_R$ . Model 4 provides I/M ratios that are specific to each district and these ratios are then applied to the district-specific mortality to predict the incidence. The

derivations of the  $I_R$  estimates using Model 2 and Model 4, which are compared here, are detailed in Appendix A3.

## Implementation

All analyses were performed in R, version 3.4.3, using the *gam* function from the package *mcgv*, version 1.8–23.<sup>3</sup> The restricted maximum-likelihood criterion was used to estimate the smoothing parameters<sup>3</sup>(p.262).

## Results of the validation assessment

### First assumption: national incidence may be currently estimated from the incidence data of registries

Regarding the first assumption, let us first illustrate the way the reference estimates are obtained using the HA indicator. Table 1 shows the observed number of HA cases and incident cases in the registry area over the period 2011–2015. The observed number of HA cases overestimates the number of incident cases with the ratio HA/I = 1.05 for female lung cancer and 1.07 for prostate cancer. As shown in Supplementary Figure 2, available as Supplementary data at *IJE* online, this ratio varies only slightly with age in female lung cancer, but varies from 0.8 to 1.5 at age 90 for prostate cancer, which means that, at age 90, there are 1.5 more HA cases than incident cases. Applying these ratios by age to the numbers of national HA cases by age leads to annual estimates of 11 792 female lung-cancer-incidence cases and 47 769 prostate-cancer-incidence cases (see details in Appendix A2).

Table 2 presents the number of cases estimated in France over the period 2011–2015 from Model 2 compared with the three reference estimates (H/I, A/I or HA/I method), which is the key validation element. The proposed method is labelled the ‘new’ method in this table. In addition to lung- and prostate-cancer examples, the site ‘all cancers’ is also presented here. Relative differences were small, at around –4% for lung cancer in women, +4% for prostate cancer and <2% for the ‘all cancers’ site. Supplementary Table 3, available as Supplementary data at *IJE* online, presents these results for the 22 cancer sites studied and Supplementary Table 4, available as

Supplementary data at *IJE* online, shows the corresponding age-standardized incidence rates with 95% confidence interval (CI). Overall, the relative differences between the new and the reference estimates were small. The absolute values of these relative differences as compared with the mean reference were <5% except for testis cancer in men (12%) and for larynx cancer (17%, although numbers are very small), kidney cancer (7%) and stomach cancer (6%) in women. Details by age may be found in Supplementary Figure 3, available as Supplementary data at *IJE* online, and the differences did not exhibit strong age patterns.

To examine whether districts with a registry may be seen as a random sample from French districts in terms of incidence rates, Supplementary Figure 4, available as Supplementary data at *IJE* online, presents the cancer HA rate by district in increasing order, those with a cancer registry being indicated in red. Except for testis cancer and larynx cancer in women, this figure shows that the assumption is reasonable and is reassuring regarding the variance accuracy.

### Second assumption: incidence trends are similar in the registry area and in the whole of France

The second assumption was indirectly examined by comparing the mortality trends. Supplementary Figure 5, available as Supplementary data at *IJE* online, shows that the trends in age-standardized mortality rates in the registry area and the whole of France are overall similar and often the two curves coincide.

### Third assumption: incidence trends in the registry area may be correctly estimated despite incomplete histories

Supplementary Figure 6, available as Supplementary data at *IJE* online, shows the incidence trends in the registry area from 1985 to 2015 (age-standardized rates) as estimated by the proposed method and by the alternative method that uses mortality. Before 1990 (dotted vertical line), some differences were observed for a few cancer sites; afterwards, the trends were similar for all cancer sites, suggesting that Model 2 may be used to estimate the trends from 1990 despite incomplete histories.

**Table 1** Use of HC/I ratio to estimate national incidence (reference estimates): illustration with HA indicator, France, 2011–2015—female lung cancer and prostate cancers (annual numbers)

Cancer site	Registry area			France	
	Observed HA cases	Observed incident cases	All age HA/I ratio	Observed HA cases	Estimated incident cases <sup>a</sup>
Lung, female	2342	2223	1.05	12 396	11 792
Prostate	10 587	9881	1.07	51 316	47 769

<sup>a</sup>National estimation obtained with Model 3 (see Appendix A2).



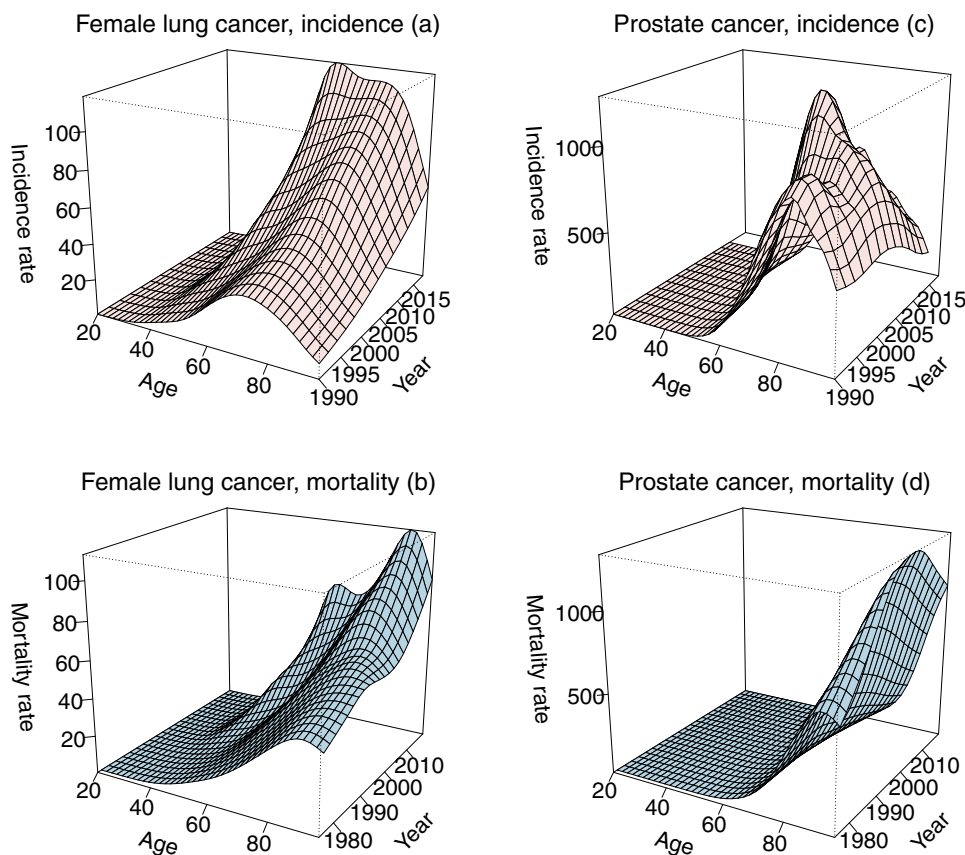
**Table 2** New method compared with the reference (HA/I, A/I or H/I): estimated annual numbers of cancer cases and relative differences, France, 2011–2015—female lung cancer, prostate cancer and ‘all cancers’

Cancer site	Estimated annual numbers of cases				Relative differences (%) <sup>a</sup>				
	New	HA/I	A/I	H/I	Mean <sup>b</sup>	HA/I	A/I	H/I	Mean <sup>c</sup>
Lung, female	11 251	11 792	11 582	11 873	11 749	–5	–3	–5	–4
Prostate	50 030	47 769	47 310	48 624	47 901	5	6	3	4
All cancers, male	201 293	198 061	194 218	198 413	196 897	2	4	1	2
All cancers, female	163 319	163 878	162 704	163 327	163 303	0	0	0	0

<sup>a</sup>Relative difference = 100\*(new-reference)/reference.

<sup>b</sup>Mean of the three reference estimates HA/I, A/I and H/I.

<sup>c</sup>Relative difference as compared with the mean of three reference estimates HA/I, A/I and H/I.



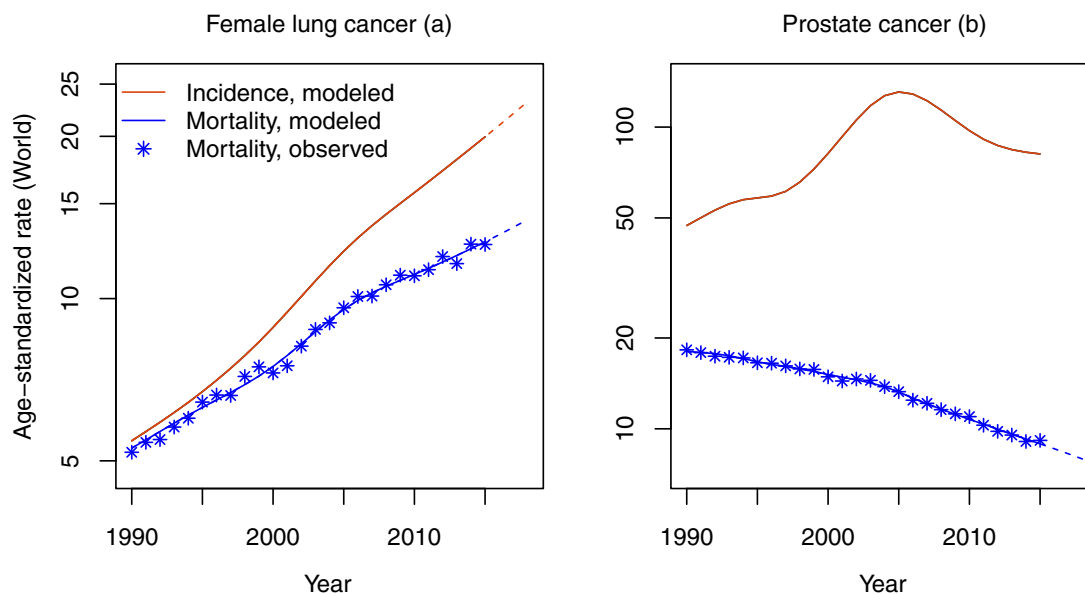
**Figure 1** 3D plot of national incidence (1990–2018<sup>a</sup>) and mortality (1975–2018) rates by age and year, France, female lung cancer and prostate cancer. <sup>a</sup>1990–2015 for prostate-cancer incidence.

**Application**

The application presents detailed national incidence and mortality-trend analyses for female lung cancer and prostate cancer in France (1990–2018). The rates are expressed per 100 000 person-years and the reference population used for age-standardization is the world population.<sup>16</sup>

Figure 1 shows the national mortality and incidence rates by age and year as obtained from Model 1 and Model 2, respectively. This figure illustrates that MPS

provide smooth surfaces in both directions of age and year, and may model simple surfaces (e.g. female lung-cancer incidence) as well as complex surfaces (e.g. prostate-cancer incidence). The goodness of fit of such models is shown in Supplementary Figure 7, available as Supplementary data at *IJE* online, for incidence (prostate-cancer example) and Figures 8, available as Supplementary data at *IJE* online, for mortality (female lung-cancer example). Figure 1 was obtained from predicted rates by annual age and year, but,



**Figure 2** Trends in incidence and mortality age-standardized rates (log-scale), France, 1990–2018\*. Female lung cancer (a) and prostate cancer (b). \*1990–2015 for prostate-cancer incidence.

obviously, as a birth cohort equals the year minus the age, the predicted rates are actually available by age, year and cohort. These rates may then be summarized by age-standardized rates (Figure 2) or described precisely by cross-sectional cuts of Figure 1 in the age, year or cohort direction (Figures 3 and 4).

Figure 2 shows the trends in age-standardized incidence and mortality rates from 1990 to 2018 (see also Supplementary Table 5, available as Supplementary data at *IJE* online). Both indicators increased dramatically in female lung cancer, especially the incidence. For prostate cancer, mortality decreased, especially in recent years, whereas the incidence increased dramatically up to year 2005, then declined.

This synthetic picture may be refined by looking at the trends in the rates by age, year or cohort (Figures 3 and 4). For female lung cancer (Figure 3), the incidence and mortality-trend patterns were overall consistent. Figure 3a and b shows the incidence and mortality rates by age for different years (*transversal* age curves); although this representation is very common, these curves do not represent properly the *lifetime ageing* effect. For instance, the incidence age curve for the year 2015 shows similar rates in women aged 60 and 80; however, this curve does not compare risks at different ages of the same women (the former were born in 1955 and the latter in 1935). A proper representation of lifetime experience is to plot rates by age for different birth cohorts (*longitudinal* age curves) as in Figure 3c and d. These are key figures to describe the progress of cancer risk over life for a given cohort and the way

this risk has evolved over successive cohorts (i.e. trends). Figure 3c and d show that, for female lung cancer, whatever the birth cohort, the incidence rates and mortality rates increased with age over the lifetime (including ages 60–80, contrarily to what Figure 3a may suggest). Furthermore, these figures show that these rates increased considerably over successive cohorts. To complement these key figures and focus on the trends, the rates may be plotted by year at different ages (Figure 3e and f) or, more advantageously, by cohort (Figure 3g and h) using log-scales for easier interpretation. The latter figures show a marked increase in the incidence and mortality rates between the cohorts from 1940 to nearly 1960, followed by a slowdown or even stabilization for women born after that.

The incidence and mortality trends diverged for prostate cancer (Figure 4). The mortality decreased regularly at all ages (Figure 4f and h) with a slightly more pronounced decrease in older ages and recent years. Contrariwise, the incidence trends were especially complex. Figure 4a shows marked changes in transversal age curves, with the 2005 age curve standing high above all others; however, the underlying phenomena are not easy to understand from this figure. Once again, Figure 4c makes the picture clearer: the longitudinal age curves evolved strongly with birth cohorts; there was a nearly 10-year shift towards younger ages between the successive cohorts of 1920, 1930 and 1940. In these cohorts, the incidence peaked at ages 82, 74 and 66, respectively (and thus these peaks occurred in the years 2002, 2004 and 2006). In Figure 4c, the period

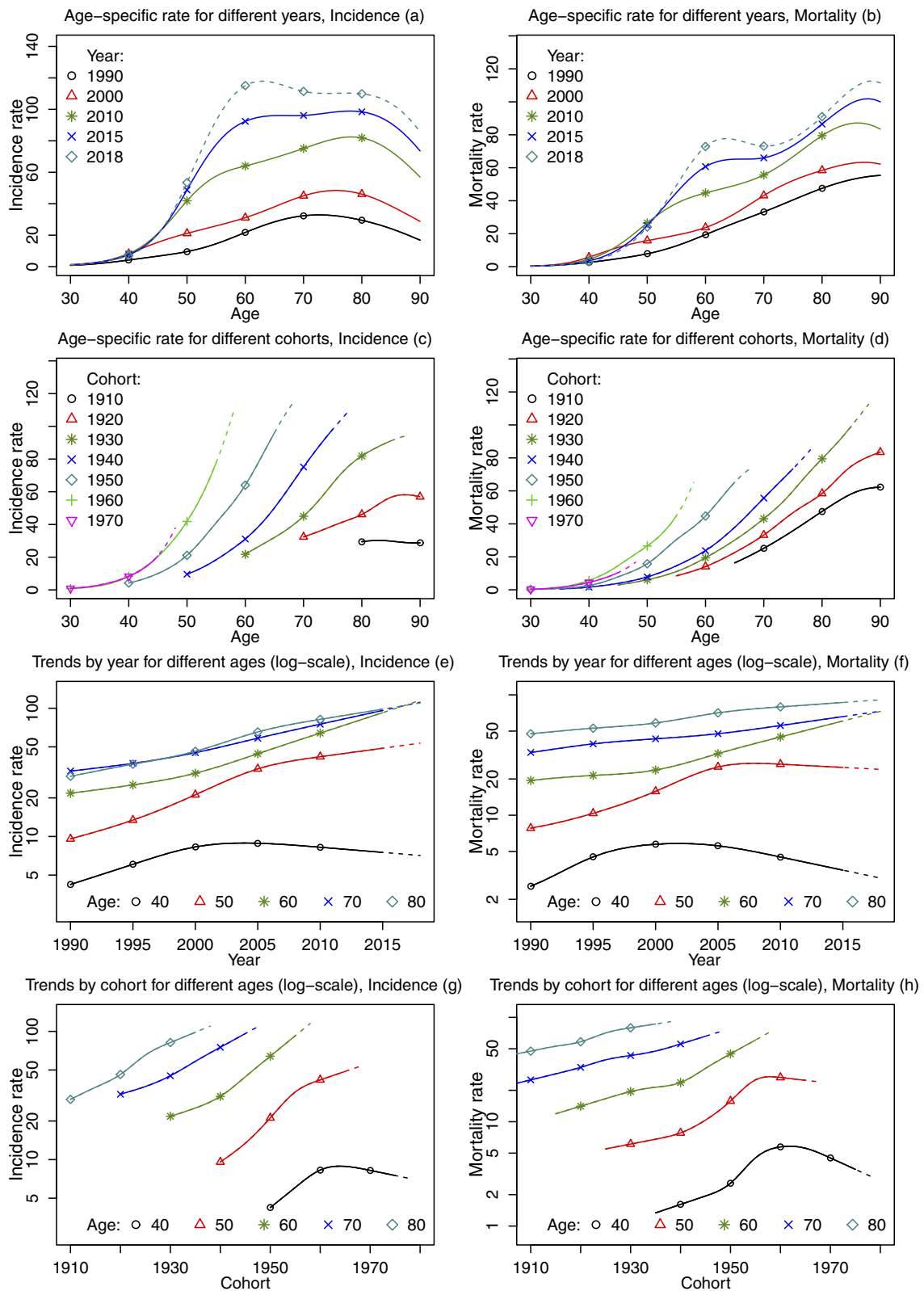
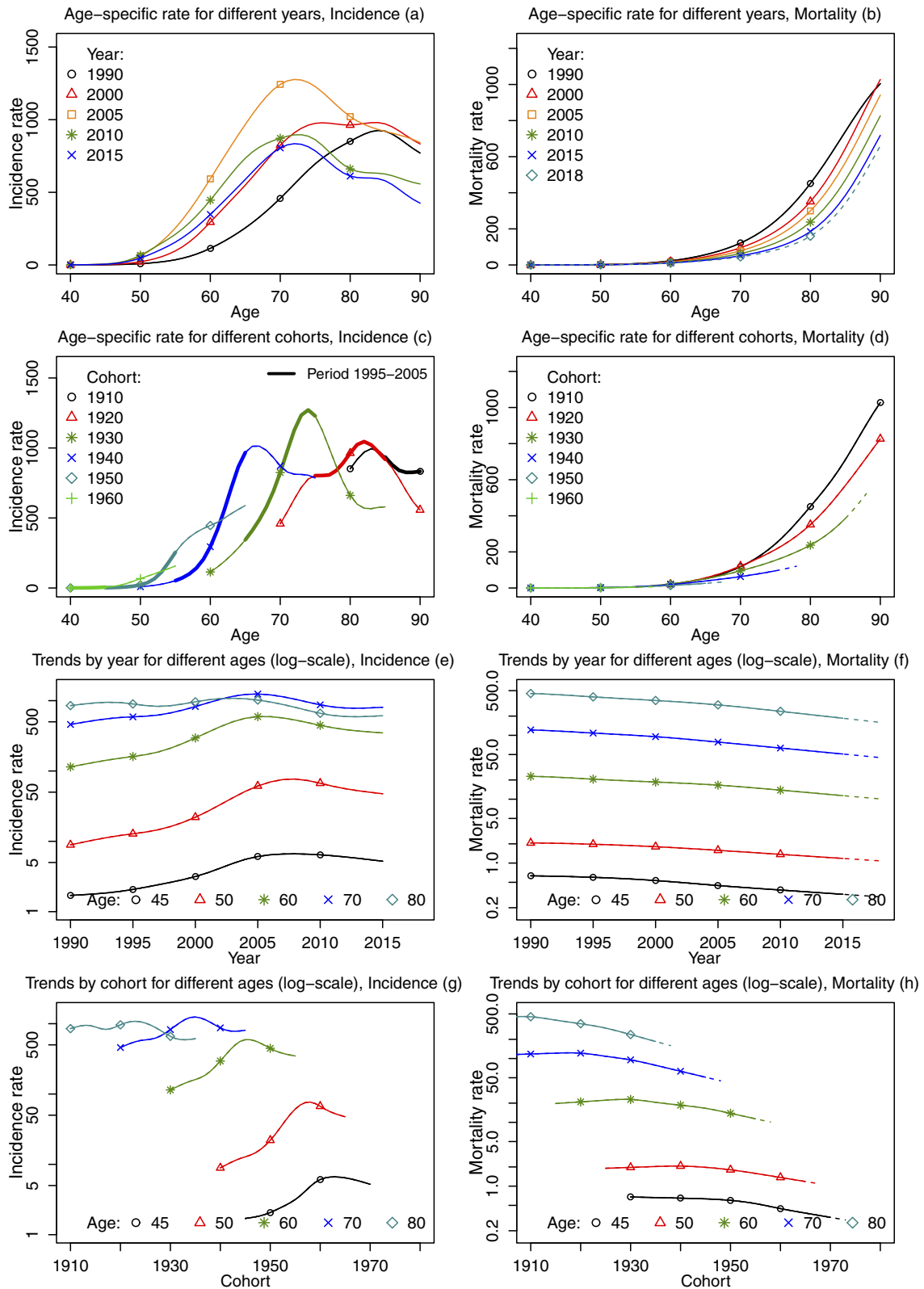


Figure 3 Trends in incidence and mortality rates by age, year and birth cohort, female lung cancer, France, 1990–2018.





**Figure 4** Trends in incidence and mortality rates by age, year and birth cohort, prostate cancer, France, 1990–2018<sup>a</sup>.  
<sup>a</sup>1990–2015 for prostate-cancer incidence.

1995–2005 over which prostate-specific antigen (PSA) screening particularly increased in France is highlighted. The trends by year (Figure 4e) complement this picture and show that the incidence increased dramatically up to 2005 (especially since 1998) in all ages, except the oldest, and then declined; this decline has slowed down, though, in the most recent years for ages 70–80.

## Discussion

### Validation of the French national incidence estimates

We proposed an approach to estimate the national cancer-incidence trends using only local-registry data. The validity assessment confirmed the acceptability of the main assumptions in France. The results demonstrate that it is possible to obtain valid national incidence estimates with a registry area that represents only 20% of the population, without the need for correction. The key condition is that the district cancer incidence should have the same mean and variability within the registry area and the whole country. In addition, trends could be estimated since 1990 despite incomplete histories in some registries. Because it does not use mortality, this methodology could also be applied to estimate the incidence for 22 malignant hemopathies and for various histological subtypes.<sup>17</sup>

Validation studies for national incidence estimations are scarce because there is no gold standard. An extensive external validation of the I/M method previously used in France was carried out for 22 cancer sites using HC data.<sup>13</sup> The accuracy of these I/M estimates was similar to those obtained here with the proposed method. Recently, Antoni *et al.*<sup>18</sup> assessed the various methods used in GLOBOCAN to estimate the future national incidence of 25 cancer sites taking the Norway incidence as the gold standard and mimicking a subnational registry area. However, this validation process is not suitable in a country without a nationwide cancer registry. We proposed here a comprehensive validation process, using extensively external data to validate recent national estimates and their trends. In the era of big data, the availability of HC data has increased considerably over the last decade, offering the possibility of external validation in many countries.

### Suitability of MPS for incidence and mortality-trend analyses

MPS and more generally penalized-regression splines are still rarely used for trend analyses, although an unpenalized version was proposed 30 years ago<sup>12</sup> and then various penalized versions for projecting all-cause mortality or cancer incidence or mortality.<sup>19,20</sup>

Yet, MPS present several advantages for trend analyses, as illustrated on the lung- and prostate-cancer examples (Figures 1–4): (i) they allow age and year to be analysed as continuous variables, which avoids the loss of information due to categorization; (ii) they account for non-linearity and age–year interactions, as exemplified in the analyses of prostate-cancer incidence in which MPS managed to catch the complex age–year interaction; (iii) they may model simple as well as complex trends, the smoothing parameter estimation acting as a model-selection procedure; (iv) they provide smooth estimates according to year and age, and thus also birth cohort<sup>21</sup>; (v) the MPS used here, being parametric-regression splines, rather than smoothing splines,<sup>22</sup> allow direct derivation of predictions and CIs from the parameters and their variance–covariance matrix.

When using MPS, only the marginal bases have to be specified. Restricted cubic splines is an usual choice, although other relevant choices are possible, e.g. thin-plate splines<sup>3</sup>; in addition, restricted splines are well adapted for short projections, since they are constrained to be linear beyond the boundary knots. Other types of penalized splines such as the P-splines proposed by Eilers and Marx could also be used.<sup>23</sup> A general principle when using MPS is to choose a slightly higher number of knots for the splines than the number deemed necessary and let penalization avoid overfitting. The choices regarding the knots (number and location) are thus less critical in the penalized than in the unpenalized framework.<sup>3,24</sup> A detailed discussion of these aspects may be found in these two references. In the case of projection, the last knot for  $g(y)$  may be located 5 years before the last observation, as done here, to stabilize the projection.

In the literature, age-period-cohort (APC) models are often used for incidence and mortality-trend analyses, with variables treated as qualitative or continuous variables.<sup>12,21,25–27</sup> APC models attempt to decompose  $\text{Log}(\lambda_{a,y,[c]})$  as  $f(a) + g(y) + h(c)$  ( $c$  being the birth cohort) but, as  $p = a + c$ , this decomposition raises identifiability and interpretation issues.<sup>12,28</sup> Here, we were interested in getting smooth estimates of this rate rather than decomposing it. In this ‘smoothing’ perspective, APC models using univariate penalized splines may be also implemented with the R package *mgcv* (the non-identifiability being handled in an internal procedure,<sup>3</sup> p. 233). Penalized APC models share with MPS models the advantages of penalization and are reduced to age-cohort (AC) or age-period (AP) models in case of strong penalization, which is an interesting feature. However, as illustrated by the prostate-incidence analysis shown in Supplementary Figure 7, available as Supplementary data at *IJE* online, the APC may remain too constrained to fit very complex trends and, in such cases, it is difficult to figure out which constraints are responsible for the lack of fit (e.g. underestimation at ages

>80 in 2015). Furthermore, we also preferred MPS to penalized APC models because they are identifiable and they conform to the fact that the incidence or mortality rates only have two dimensions and not three (cohort equals year minus age), which is a major advantage in our view. Nevertheless, a comprehensive simulation-based study comparing MPS and penalized APC models would be interesting. Note that we used an MPS model of age and year rather than of age and cohort, because the  $(a, y)$  surface is fully observed whereas the  $(a, c)$  surface is only observed on a diagonal band, this unbalanced design being obviously less favourable for the estimation procedure.

From a more general perspective, MPS constitute a general regression approach that is convenient for many contexts, including survival-trend analyses,<sup>29</sup> spatio-temporal smoothing<sup>30</sup> or modelling of seasonal phenomena using a cyclic basis<sup>3</sup> (p. 371).

### Epidemiological interest of detailed trend analyses

The two cancer sites presented here contrast highly in terms of epidemiological pattern and context; yet, an MPS model allowed the trends for both sites to be described precisely (Figures 3 and 4). This information may be linked to the dynamics of known or suspected risk factors, supports hypotheses about factor contribution and allows refined epidemiological interpretations.

A dramatic increase in female lung-cancer incidence and mortality was observed, especially among women born from 1940 to ~1960; this increase slowed down among women born afterwards (Figure 3g and h). This dramatic increase may be strongly linked with the important rise in tobacco consumption by women since the 1950s, which massively involved women born from 1940 to 1960.<sup>31</sup> In light of the present findings, it is likely that this trend will continue among these generations of heavy smokers as women get older. The slowdown in cohorts born after 1960 (observed up to age 55) may be related to the slowdown observed in tobacco consumption since 2000 in women aged 40–44 (see Figure 2 from Hill *et al.*<sup>31</sup>). Furthermore, as expected for such a lethal cancer as lung cancer, the trends in incidence and mortality were very similar, though the results stem from completely separate data; this shows the reliability of MPS.

Regarding prostate cancer, the mortality decreased whereas the incidence increased markedly up to 2005 before declining. The detailed incidence trends were complex but easier to figure out by looking at the longitudinal age curves in successive cohorts (Figure 4c). Although other risk factors may have contributed to the marked changes observed, the massive development of PSA screening is probably largely responsible for these evolutions.<sup>32</sup> Indeed,

men have been offered PSA screening since the late 1990s or early 2000s in France, so all cohorts were impacted from then onwards. In the short term, screening increases incidence and advances both the year and the age at diagnosis; in the long term, screening removes previously early-detected cancers from upcoming incidence. These phenomena are clearly illustrated by Figure 4c, which shows an important age shift in the longitudinal age curves towards younger ages; for cohorts from 1920 to 1940, these curves reached a peak before a sharp decline. These declines started in around 2005 and result probably from the long-term effects of former screenings plus a stability or decrease in screening practices.<sup>33</sup> Regarding mortality, the observed decrease resulted probably from a decrease in the incidence of unfavourable cases due to early detection combined with the improvement in cancer treatment.

### Conclusion

The proposed method provided valid national incidence-trend estimates in France. The validation process may be carried out in any country where exhaustive HC data are available and may help in choosing the most appropriate national estimation method for each country's specific context.

Incidence and mortality-trend analyses for lung and prostate cancers illustrated the suitability of MPS for such analyses and the epidemiological interest of providing detailed results by age and year or age and birth cohort.

MPS form a powerful statistical tool for incidence, mortality and survival-trend analyses. For those not familiar with penalization, some investment is needed to become comfortable with its theory and practice; afterwards, MPS and, more generally, penalized-regression splines will prove easy to implement and very useful.

### Supplementary data

Supplementary data are available at *IJE* online.

### Funding

This work was supported by the Institut National du Cancer (INCa), grandt N°2016-192, and by Santé Publique France (SPF).

### Acknowledgements

The authors thank Jean Iwaz for his thorough proofreading.

### Author contributions

Z.U., E.D., E.C., L.Ro., M.F., M.C., N.B. and L.Re. conceived the overall design and the statistical methods. Z.U., E.C. and E.D. implemented the statistical analyses. Z.U. reviewed the literature and wrote the first draft of the manuscript, supervised by L.Re. G.D., P.G., M.C., S.P.L. and A.M. interpreted the epidemiological results and were involved in the final decision of adopting the proposed method. All authors reviewed the manuscript.

## Conflict of interest

None declared.

## References

1. Ferlay J, Colombet M, Soerjomataram I *et al.* Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer* 2019;144:1941–53.
2. Wood SN, Augustin NH. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol Modell* 2002;157:157–77.
3. Wood SN, *Generalized Additive Models: An Introduction with R*, 2nd edn. London: Chapman & Hall/CRC, 2017.
4. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *J Am Stat Assoc* 2016;111:1548–75.
5. Galceran J, Ameijide A, Carulla M *et al.* Cancer incidence in Spain, 2015. *Clin Transl Oncol* 2017;19:799–825.
6. Belot A, Grosclaude P, Bossard N *et al.* Cancer incidence and mortality in France over the period 1980–2005. *Rev Epidemiol Sante Publique* 2008;56:159–75.
7. Remontet L, Esteve J, Bouvier AM *et al.* Cancer incidence and mortality in France over the period 1978–2000. *Rev Epidemiol Sante Publique* 2003;51:3–30.
8. Yang L, Parkin DM, Ferlay J, Li L, Chen Y. Estimates of cancer incidence in China for 2000 and projections for 2005. *Cancer Epidemiol Biomarkers Prev* 2005;14:243–50.
9. Binder-Foucard F, Bossard N, Delafosse P *et al.* Cancer incidence and mortality in France over the 1980–2012 period: solid tumors. *Rev Epidemiol Sante Publique* 2014;62:95–108.
10. Uhry Z, Belot A, Colonna M *et al.* National cancer incidence is estimated using the incidence/mortality ratio in countries with local incidence data: is this estimation correct? *Cancer Epidemiol* 2013;37:270–77.
11. Chatignoux E, Remontet L, Iwaz J, Colonna M, Uhry Z. For a sound use of health care data in epidemiology: evaluation of a calibration model for count data with application to prediction of cancer incidence in areas without cancer registry. *Biostatistics* 2019;20:452–67.
12. Heuer C. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics* 1997;53:161–77.
13. Uhry Z, Remontet L, Colonna M *et al.* Cancer incidence estimation at a district level without a national registry: a validation study for 24 cancer sites using French health insurance and registry data. *Cancer Epidemiol* 2013;37:99–114.
14. Colonna M, Mitton N, Remontet L *et al.* Méthode d'estimation de l'incidence régionale des cancers à partir des données d'incidence des registres, des données de mortalité par cancer et des bases de données médico-administratives. *Bull Epidemiol Hebd* 2013;43:566–74.
15. Chatignoux E, Remontet L, Colonna M, Grosclaude P, Decool E, Uhry Z. Estimations régionales et départementales d'incidence et de mortalité par cancers en France, 2007–2016: évaluation de l'utilisation des données médico-administratives pour estimer l'incidence départementale: comparaison de l'incidence observée et prédite dans les registres sur la période 2007–2014. Saint-Maurice: Santé Publique France, 2019. <http://invs.santepubliquefrance.fr/Dossiers-thematiques/Maladies-chroniques-et-traumatismes/Cancers/Donnees-par-territoire> (27 April 2020, date last accessed).
16. Doll R, Payne P, Waterhouse J. *Cancer Incidence in Five Continents, Vol. I: A Technical Report*. Berlin: Springer-Verlag, 1966.
17. Defossez G, Le Guyader-Peyrou S, Uhry Z, *et al.* *Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018, Vol. 1: Tumeurs solides: Étude à partir des registres des cancers du réseau Francim*. Saint-Maurice: Santé Publique France, 2019. <https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-sein/documents/rapport-synthese/estimations-nationales-de-l-incidence-et-de-la-mortalite-par-cancer-en-france-metropolitaine-entre-1990-et-2018-volume-1-tumeurs-solides-etud> (27 April 2020, date last accessed).
18. Antoni S, Soerjomataram I, Moller B, Bray F, Ferlay J. An assessment of GLOBOCAN methods for deriving national estimates of cancer incidence. *Bull World Health Organ* 2016;94:174–84.
19. Clements MS, Armstrong BK, Moolgavkar SH. Lung cancer rate predictions using generalized additive models. *Biostatistics* 2005;6:576–89.
20. Katanoda K, Kamo K, Saika K *et al.* Short-term projection of cancer incidence in Japan using an age-period interaction model with spline smoothing. *Jpn J Clin Oncol* 2014;44:36–41.
21. Carstensen B. Age-period-cohort models for the Lexis diagram. *Stat Med* 2007;26:3018–45.
22. Hastie T, Tibshirani R. *Generalized Additive Models*. London: Chapman and Hall, 1990.
23. Eilers PH, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996;11:89–121.
24. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. New York: Cambridge University Press, 2003.
25. Clayton D, Schifflers E. Models for temporal variation in cancer rates: II: Age-period-cohort models. *Stat Med* 1987;6:469–81.
26. Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics* 1983;39:311–24.
27. Smith T, Wakefield J. A review and comparison of age-period-cohort models for cancer incidence. *Stat Sci* 2016;31:591–610.
28. Wilmoth JR. Variation in vital rates by age, period, and cohort. *Sociol Methodol* 1990;20:295–335.
29. Remontet L, Uhry Z, Bossard N *et al.* Flexible and structured survival model for a simultaneous estimation of non-linear and non-proportional effects and complex interactions between continuous variables: performance of this multidimensional penalized spline approach in net survival trend analysis. *Stat Methods Med Res* 2019;28:2368–84.
30. Ugarte MD, Goicoa T, Etxeberria J, Militino AF. Projections of cancer mortality risks using spatio-temporal P-spline models. *Stat Methods Med Res* 2012;21:545–60.
31. Hill C, Jouglu E, Beck F. Le point sur l'épidémie du cancer du poumon dû au tabagisme. *Bull Epidemiol Hebd* 2010;19-20:210–13.
32. Grosclaude P, Remontet L, Marliac Velten DL, Uhry M, Leone Z. Le cancer de la prostate: incidence, survie et mortalité en France. *Bull Epidemiol Hebd* 2016;39-40:693–99.
33. Tuppin P, Leboucher C, Doug M, Peyre-Lanquar G, Gabach P, Descotes J. Dépistage individuel du cancer de la prostate chez les hommes de 40 ans et plus, France. Données du Système National D'information Inter-Régimes de L'Assurance Maladie. *Bull Epidemiol Hebd*. 2016;39-40:700–6.

## Appendix

### A1. Construction of a tensor product from the marginal bases

The tensor product  $te(a, y)$  of the marginal bases  $f = (f_m)_{m=1..M}$  of age and  $g = (g_l)_{l=1..L}$  of year is obtained by a multiplication term by a term of the two bases:

$$te(a, y) = \sum_m \sum_l \beta_{l,m} \cdot f_m(a) \cdot g_l(y),$$

where

$(\beta_{l,m})_{l=1..L; m=1..M}$  are the  $M * L$  parameters to be estimated.

For illustration, let's consider simple quadratic bases for both age and year, and start the parameter indexation at 0 for clear interpretation; then:

$$\begin{aligned} te(a, y) = & \beta_{0,0} + \beta_{1,0} \cdot a + \beta_{2,0} \cdot a^2 + \\ & + \beta_{0,1} \cdot y + \beta_{1,1} \cdot a \cdot y + \beta_{2,1} \cdot a^2 \cdot y \\ & + \beta_{0,2} \cdot y^2 + \beta_{1,2} \cdot a \cdot y^2 + \beta_{2,2} \cdot a^2 \cdot y^2. \end{aligned}$$

### A2. Validation of the first assumption: method to estimate national incidence using the HC/I ratio, period 2011–2015

Here, we present briefly the method detailed and validated in the paper by Chatignoux *et al.*<sup>11</sup> (see <https://github.com/echatignoux/CalibInc> for R codes and tutorial). The analysis concerned data aggregated over the period 2011–2015 by district and 5-year age classes. For simplicity, we will use a unique notation  $HC$  here to refer to either the  $HA$ ,  $H$  or  $A$  indicator. The following model was used,  $a_i$  being the central age of age class  $i$ :

$$\begin{aligned} HC_{j,i} | K_{j,i} & \sim \text{Poisson}(\rho_{j,i} \cdot K_{j,i}) \text{ and} \\ \text{Log}(\rho_{j,i}) & = s(a_i) + b_j, \text{ with } b_j \sim N(0, \sigma_{HC,I}^2), \end{aligned}$$

Model 3

where  $\text{Log}(K_{j,i})$  is an on offset in the model,  $\rho$  represents the HC/I ratio,  $s$  is a thin-plate spline of age with as many knots as the number of age classes and  $b_j$  is a district random effect.

The reference national incidence estimations for age class  $i$  are then derived as:

$$\hat{\lambda}_i^{FR} = \frac{\hat{K}_i^{FR}}{PA_i^{FR}}, \text{ with } \hat{K}_i^{FR} = H_i^{FR} / \exp(\hat{s}(a_i) + \sigma_{HC,I}^2/2),$$

### A3. Validation of the third assumption: alternative method to estimate incidence in the registry area from 1990 to 2018 using mortality

For an alternative estimate of the incidence trend in the registry area, the I/M ratio was modelled using data from all registries from 1985 to 2015, aggregated by year  $y$ , district  $j$  and 5-year age classes  $i$  (centred on age  $a_i$ ). Mortality was previously and separately smoothed in each district using a tensor of age and year. The model is:

$$\begin{aligned} K_{j,i,y} | \hat{D}_{j,i,y} & \sim \text{Poisson}(\gamma_{j,i,y} \hat{D}_{j,i,y}) \text{ and} \\ \text{Log}(\gamma_{j,i,y}) & = te_\gamma(a_i, y) + \nu_j, \text{ with } \nu_j \sim N(0, \sigma_{I,M}^2). \end{aligned}$$

Model 4

In this model,  $\hat{D}_{j,i,y}$  is the estimated number of deaths from preliminary smoothing and  $\text{Log}(\hat{D}_{j,i,y})$  is an offset,  $\gamma$  is the I/M ratio,  $te_\gamma(a_i, y)$  a tensor of age and year, and  $\nu_j$  a district random effect. For this analysis, except for nervous central-system cancers, ages  $\leq 20$  were excluded to avoid modelling the I/M ratio where the incidence and mortality are almost zero and the ratio I/M is hardly defined (or ages  $\leq 15$  for testis cancer, thyroid cancer and Hodgkin lymphoma). The district effect was not entered into the model for thyroid cancer and testis cancer that had mortality rates  $< 0.5$  per 100 000 person-years.

Derivation of incidence in the registry area with the proposed method (from Model 2) and with an alternative method using mortality (from Model 4)

The incidence in the registry area was estimated by summing the estimates of all districts:

$$\hat{K}_{i,y}^R = \left( \sum_j \hat{K}_{j,i,y} \right),$$

where  $\hat{K}_{a,i,y} = \exp(\hat{te}_\lambda(a, y) + \hat{u}_j) \cdot PY_{j,a,y}$  for the proposed method, using parameters from Model 2, or  $\hat{K}_{j,i,y} = \exp(\hat{te}_\gamma(a_i, y) + \hat{\nu}_j) \cdot \hat{D}_{j,i,y}$  for the alternative method, using parameters from Model 4.