Research Article

Statistics in Medicine

Received 27 June 2013, Accepted 26 May 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6240

Estimating cancer incidence using a Bayesian back-calculation approach

Leonardo Ventura[a*†] and Maura Mezzetti[b]

We propose a Bayesian hierarchical model for the calculation of incidence counts from mortality data by a convolution equation that expresses mortality through its relationship with incidence and the survival probability density. The basic idea is to use mortality data together with an estimate of the survival distribution from cancer incidence to cancer mortality to reconstruct the numbers of individuals who constitute previously incident cases that give rise to the observed pattern of cancer mortality. This model is novel because it takes into account the uncertainty from the survival distribution; thus, a Bayesian-mixture cure model for survival is introduced. Furthermore, projections are obtained starting from a Bayesian age-period-cohort model. The main advantage of the proposed approach is its consideration of the three components of the model: the convolution equation, the survival mixture cure model and the age-period-cohort projection within a directed acyclic graph model. Furthermore, the estimation are obtained through the Gibbs sampler. We applied the model to cases of women with stomach cancer using six age classes [15–45], [45–55], [55–65], [65–75], [75–85] and [85–95] and validated it by using data from the Tuscany Cancer Registry. The model proposed and the program implemented are convenient because they allow different cancer disease to be analysed because the survival time is modelled by flexible distributions that are able to describe different trends. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: incidence; mortality; survival; Bayesian hierarchical model; cure model

1. Introduction

Measuring the burden of disease is one of the main concerns of public health. Related measures are necessary to describe the general state of population's health, to establish public health goals and to compare the national health status and performance of health systems across countries. Furthermore, such studies are needed to assess the allocation of health care and health research resources across disease categories and to evaluate the potential costs and benefits of public health interventions [1].

Cancer is a major concern worldwide, as it decreases the quality of life and leads to premature mortality. Indeed, the study and estimation of the cancer epidemic as a whole, with respect to its dimension, trends and projections, is essential. Incidence, prevalence, survival and mortality are the most important indicators in measuring the burden of cancer. These indicators refers to four different aspects of the same morbidity process, and simultaneous monitoring of the four indicators is required to understand and fully describe different aspects of the disease. However, complete information on all of these variables is seldom available. While mortality data are available from official statistics, incidence and prevalence rates, in general, can only be obtained from disease registries. Such data are usually rarely available at the national or regional levels. In particular, mortality data have been available in Italy since 1970, while incidence and survival data have recently become available from local population-based cancer registries covering approximately 34% of the Italian population [2].

The idea for this paper arose from the necessity of estimating cancer incidence rates. Mortality alone is not able to provide detailed estimates of the cancer burden; for cancers with good prognosis, it is simply not suitable to study the associations between risk factor changes and cancer occurrence. Improvements in cancer patients' survival as well as the possibility of cures for some forms of cancer has led the incidence

[a] *Cancer Prevention and Research Institute, Florence, Italy*
[b] *Department Economics and Finance, Università Tor Vergata, Rome, Italy*
**Correspondence to: Leonardo Ventura, Cancer Prevention and Research Institute, Florence, Italy.*
[†] *E-mail: l.ventura@ispo.toscana.it*

and mortality rates to diverge. Furthermore, knowledge of the incidence is necessary to evaluate screening efficacy and to plan new screening therapies.

Numerous different statistical methods have been developed to estimate population-based cancer incidence in areas not covered by cancer registries. The family of methods based on the incidence-to-mortality ratio [3, 4] is the most widely used in estimating incidence rates. The other most important methods to estimate cancer incidence obtain the latter quantity by back-calculating from mortality and survival data. A deterministic back-calculation approach of incident and prevalent cases was developed by Verdecchia *et al* [5]. This method, which was developed along with a computer package named Mortality Incidence Analysis Model (MIAMOD) [6] and has been widely applied to provide cancer incidence estimations both in Europe and the U.S. [7–10], is based on transitional equations linking cancer-specific mortality and prevalence to the probabilities of being diagnosed with cancer (incidence) and of surviving after such a diagnosis (relative survival). Cancer incidence is modelled using an age-period-cohort (APC) model. Cancer-specific mortality and relative survival are assumed to be known input data.

A back-calculation based on a Bayesian approach was previously developed by Mezzetti and Robertson [11] as a Bayesian hierarchical model that estimates the age-specific cancer incidence per year from age-specific cancer mortality. The incident cases per year are considered observations from a discrete-time stochastic process following an autoregressive structure within a Poisson regression model. The model assumes that the survival probability among those with cancer is known. Age effect on incidence, survival and mortality is introduced, but the uncertainty regarding the input data (survival) is not taken into account.

A crucial point in back-calculation methods has always been the survival function. To obtain reliable incidence estimates, the survival time has to be properly modelled to smooth and project the observed data and to take the relative uncertainty into consideration. Methods for survival modelling have improved drastically over recent decades [12].

Cure models are survival models that were developed to estimate the proportion of patients cured in a clinical trial. These models estimate the cured proportion and also the probability of survival of the uncured patients up to a given point in time. The model developed by Boag [13] estimates the proportion of patients cured out of those who received treatment at different cancer sites. This model is called the mixture model, because it can estimate the proportion of patients cured and the survival function of the uncured patients. Boag [13] modelled the survival function of the failure time of uncured patients by a product of a log-normal survival function and the survival function of some background distribution for the normal population. His model was further developed by Berkson and Gage [14] and was later studied extensively by several authors. The most common parametric distributions used in a mixture models are the exponential, the Weibull, the Gompertz and the generalised F.

De Angelis *et al.* [15] proposed a parametric mixture relative survival model that has been applied to individual data on colon cancer. A logistic function and the exponential or Weibull densities can be combined to model the proportion of cured patients and the failure time distribution for fatal patients, respectively. The parameters concerned are estimated by the means of the maximum likelihood method. Age, gender, period of diagnosis, stage and subsite are included in the model parameters as explicative covariates. Seppä *et al.* [16] applied a mixture cure fraction model with random effects to the cause-specific survival data of female breast cancer patients collected by the population-based Finnish Cancer Registry. In addition to the parametric mixture cure models, there are a number of nonparametric and semiparametric mixture cure models in the literature [17–19]. More recently, Yu *et al.* [20] proposed a Bayesian approach to a parametric mixture cure model for population-based cancer survival data, which was extended to county level cancer survival data by including a county-specific spatial random effect. Instead of modelling the latent distribution by a fixed parametric distribution, they used a finite mixture of log-normal, log-logistic and Weibull distributions. The parameters are estimated using the Markov chain Monte Carlo method and applied to relative survival data for colon cancer patients. The advantage of a finite mixture model based on different parametric families is the ability to accommodate the different tail heaviness of each subpopulation. A unique parametric probability density would be insufficient to correctly describe the heterogeneity of the data; the Bayesian mixture cure model that was adopted allows us to propose a more general model that is able to address survival from different types of cancer.

To project cancer incidence into the future, we first need to project the mortality counts from which the back-calculation model starts. Clayton and Schifflers [21] discussed in detail the use of APC models by using a classical approach. Several authors have proposed Bayesian APC models for this purpose. Bashir *et al.* [22] and Bray [23] used a Bayesian approach to project cancer incidence and mortality rates.

They fit an APC model to an explanatory data set and extrapolated the period and cohort effects using an autoregressive prior structure.

We propose a hierarchical Bayesian back-calculation model to estimate the cancer incidence from cancer mortality. Following the Bayesian cure model proposed by Yu *et al.* [20] without county-specific random effects, we estimate the survival function for use in the back-calculation. Furthermore, we follow the Bayesian APC model proposed by Bray [23], to project the mortality counts into the future. The three main components of our proposal, convolution equation, mixture cure survival model and APC projection, can be synthetically represented within a directed acyclic graph model. The estimation of present and future incidence rates can be obtained through Gibbs sampling. The remarkable advantage of the proposed model is its capacity to take into account the uncertainty within the estimates of survival and on mortality projection parameters in a model that is applicable to different diseases. Providing credible bounds for the final estimations and projections of incidence trends has a great impact on public health allocation strategies. Such an advantage is achieved at the cost of a higher model complexity. However, recent developments in Bayesian computation have made models of such complexity tractable justifying our approach.

The paper is organised as follows: In Section 2, our proposed Bayesian hierarchical model is shown and developed after an illustration of the most important methods proposed in the literature. In Section 3, we show the data set that is available and present our analysis. After a description of the computational aspects in Section 4, Section 5 demonstrates the application of the hierarchical model and the analysis of the data and allows for the verification of the proposed methodology and validates our results through a comparison with the observed data. A sensitivity analysis is presented in Section 6. Final conclusions and remarks are discussed in Section 7.

2. Methods

The model we propose is based on a back-calculation method using a Bayesian approach. The back-calculation method allows for the calculation of incident counts from mortality counts, through the use of a convolution equation:

$$M_t = \sum_{i=0}^{s} I_{t-i} p_{t-i,i} \tag{1}$$

where M_t is the number of deaths in year t, I_{t-i} is the number of incident cases in year $(t-i)$ and $p_{t-i,i}$ is the probability of having the failure event after exactly i years from the disease, given that the illness was in year $(t-i)$. It is assumed that there is a fixed period, of length s (in our application, 12 years), that is needed for the cancer deaths to be attributed to a specific incidence. The basic idea is to use cancer mortality data, together with an estimate of the survival distribution from cancer incidence to cancer mortality, to reconstruct the number of individuals who must have been the previous cases to give rise to the observed pattern of cancer mortality. Equation (1) becomes more complicated when age is considered,

$$M_t^{age} = \sum_{i=0}^{s} I_{t-i}^{age-i} p_{t-i,i}^{age-i} \tag{2}$$

with M_t^{age}, I_t^{age} and $p_{t-i,i}^{age}$ are now relative to people aged *age*. Equation (2) introduces a large number of parameters that are difficult to handle; moreover, unreliability in the data can be encountered. Equation (1) should be rewritten to consider age classes (in our application, [15–45], [45–55], [55–65], [65–75], [75–85] and [85–95]). Complications arise when an incident case belonging to an age class a shifts one or two age classes if it is counted as a fatal case:

$$M_t^a = \sum_{l=0}^{min(l_a,s)} \frac{l_a - l}{l_a} I_{t-l}^a p_{t-l,l}^a + \sum_{l=1}^{min(l_{a-1},s)} \frac{l}{l_{a-1}} I_{t-l}^{a-1} p_{t-l,l}^{a-1} + \sum_{l=l_{a-1}}^{s} \frac{l_{a-1} + l_a - l}{l_{a-1}} I_{t-l}^{a-1} p_{t-l,l}^{a-1} + \sum_{l=l_{a-2}}^{s} \frac{l - l_{a-1}}{l_{a-2}} I_{t-l}^{a-2} p_{t-l,l}^{a-2} \tag{3}$$

M_t^a, I_{t-l}^a and $p_{t-l,l}^a$ are now relative to the age class a; the length of that is indicated as l_a. Scalars $\frac{l_a-l}{l_a}$ and $\frac{l}{l_a}$ are introduced under the assumption that age is uniformly distributed inside each age class. The second term requires that a is a value greater than 1; the third term is present only if a is greater than 1 and s is greater than l_{a-1}; and the forth component exists only for a value for a greater that 2 and an s greater than l_{a-2}. When the equation refers to mortality counts in the last age class, the upper limit of the first summation changes to s . The previous equation can be extended in the case where s greater than $l_a + l_{a+1}$.

The basic idea behind the proposed approach is to combine three types of information in a two-stage hierarchical model. At the first stage, we assume that the number of deaths per year occurs in a Poisson distribution. We also assume a simple linear relationship between the expected mortality counts and survival rates. At the second stage, we assume an autoregressive model that acts to smooth the incidence curve and a Bayesian mixture cure model for the probability of having a fatal event after diagnosis, as shown in next subsection. The autoregressive model on the incidence counts yields estimates of a smoothed age-specific incidence curve, where the degree of smoothing is modelled through our choice of the prior distribution, without strong parametric assumptions. The previous proposed back-calculation model [11] considers the quantities $p^a_{t-i,i}$ as fixed quantities (the probability that a patient belonging to age class a has a failure event after exactly i years from the disease, given that the illness was in year $(t-i)$). This assumption, according to our conclusions, was the main drawback in the previous approach; the next paragraph shows our alternative proposal for the second level of our hierarchical Bayesian model.

2.1. Survival cure model

This subsection illustrates the Bayesian mixture cure model for the relative survival distribution, as specified in the second level of the hierarchical Bayesian model. Our model starts from the proposal of Yu *et al.* [20]. We assume access to a population-based survival data set. The advantage of population-based survival analysis is that the results of such studies are representative of the entire population, a perspective that is vital for cancer-controlling activities. It seems reasonable to assume the homogeneity of the survival data inside a homogeneous geographical area under an administrative point of view. In this case, we suppose that the health care facilities are not markedly dissimilar between the two areas.

However, in population-based cancer studies, the recorded cause of death may be not reliable because it has either been incorrectly identified or was obtained from death certificates, which are often inaccurately recorded [24]. A good way to overcome the difficulties encountered from a misspecified cause of death is to use the relative survival as a measure of survival. Relative survival is the ratio of observed (all-cause) survival to the expected survival from a comparable group in the general population. This ratio provides a measure of the excess mortality experienced by patients diagnosed with the disease of interest, regardless of whether the excess mortality is directly or indirectly attributable to the disease. The expected survival can be easily obtained from National Life Tables and is usually calculated after matching for age, gender and year of diagnosis.

Furthermore, because of improvements in cancer treatments and the dissemination of early diagnosis techniques for different types of cancer, a proportion of patients may be cured and never experience the event of interest. The cure fraction is defined as the proportion of patients who are cured of disease and become long-term survivors. Accounting for a cured fraction may improve the fit of survival data, and it may provide better predictions of long-term survival rates. The model we propose belongs to the group of survival analysis models known as cure models.

Following Yu *et al.* [20], we suppose that survival for the uncured patients can be modelled by a finite mixture of log-normal, log-logistic and Weibull distributions. The survival function of a mixture cure model is specified as

$$S(l|x) = c(x) + (1 - c(x))S_U(l|x)$$

$S_U(l|x)$ is the survival function for the uncured individuals (latent distribution), $c(x)$ is the cure fraction and x is the vector of covariates (in our application, age and period). The latent distribution $S_U(l|x)$ is assumed to be a mixture of the three parametric distributions:

$$S_U(l|x) \;=\; q_1(x) \times S_{LN}(l|x) + q_2(x) \times S_{LL}(l|x) + q_3(x) \times S_{WB}(l|x), \tag{4}$$

$$S_{LN}(l|x) \;=\; 1 - \Phi\left(\frac{logl - \mu_{LN}(x)}{\tau_{LN}}\right) \tag{5}$$

$$S_{LL}(l|x) \;=\; \left\{1 + \exp\left[\frac{logl - \mu_{LL}(x)}{\tau_{LL}}\right]\right\}^{-1} \tag{6}$$

$$S_{WB}(l|x) \; = \; \exp\left\{ -\exp\left[\frac{logl - \mu_{WB}(x)}{\tau_{WB}} \right] \right\} \tag{7}$$

Yu *et al.* [20] assume that the parameters μ_{LN}, μ_{LL} and μ_{WB} are all equal, although we prefer to assume different parameters for the three components of the mixture distribution for the uncured component; we assume the three parameters depend on the age and period.

Consider that the quantities $d_t^a(l)$ and $n_t^a(l)$ refer to patients diagnosed with a cancer of interest; that is, $d_t^a(l)$ is the number of cancer patients dying from any cause l years after the diagnosis. These patients are alive in the period t and belonging to age class a. $n_t^a(l)$ is the number of people alive in the year $t+l-1$ (among patients diagnosed with the cancer on interest), for age class a; $E_t^a(l)$ is the expected probability of surviving the interval $t+l$ for those in the general population belonging to age class a who were alive at period t. The last quantities, $E_t^a(l)$, are easily accessible from the Life Tables of the general population. The Bayesian mixture cure model proposed by Yu *et al.* [20] can be written as follows:

$$d_t^a(l) \; \sim \; Poisson\left(\lambda_t^a(l) \times n_t^a(l)\right) \tag{8}$$

$$\lambda_t^a(l) \; = \; 1 - \frac{S_t^a(l)}{S_t^a(l-1)} E_t^a(l)$$

$$S_t^a(l) \; = \; c(t,a) + (1 - c(t,a)) S_U(l|t,a) \tag{9}$$

$$S_U|t,a,q_i \; = \; q_1^{a,t} \times S_{LN}(t,a) + q_2^{a,t} \times S_{LL}(t,a) + q_3^{a,t} \times S_{WB}(t,a) \tag{10}$$

The following prior distributions are assumed:

$$\begin{aligned} q_j^{a,t} \; &= \; \frac{\exp(\nu_j)}{1 + \exp(\nu_1) + \exp(\nu_2)} \qquad j = 1,2 \\ q_1^{a,t} + q_2^{a,t} + q_3^{a,t} \; &= \; 1, \qquad \forall a, t \\ \nu_j \; &\sim \; Normal\left(0, \sigma_{\nu_j}\right) \qquad j = 1,2 \\ \mu_k \; &\sim \; Normal\left(0, \sigma_{\mu_k}\right) \\ \tau_k \; &\sim \; Gamma(0.01, 0.01) \qquad k = LN,\, LL,\, and\, WB \end{aligned}$$

Jointly with Equations (5)–(7) and recalling Equation (3), the full Bayesian hierarchical model can be written as follows:

$$M_t^a \left| I_t^a, p_{t,l}^a, I_t^{a-1}, p_{t,l}^{a-1} \right. \; \sim \; Poisson\left(\sum_{l=0}^{min(l_a,s)} \frac{l_a - l}{l_a} I_{t-l}^a p_{t-l,l}^a + \sum_{l=1}^{min(l_{a-1},s)} \frac{l}{l_{a-1}} I_{t-l}^{a-1} p_{t-l,l}^{a-1} + \ldots \right) \tag{11}$$

$$I_1^a \; \sim \; Normal\left(\mu_1^a, \sigma_1^a\right) \tag{12}$$

$$I_2^a \; \sim \; Normal\left(I_1^a, \sigma_2^a\right) \tag{13}$$

$$I_t^a \left| I_{t-1}^a, I_{t-2}^a, \sigma \right. \; \sim \; Normal\left(2I_{t-1}^a - I_{t-2}^a, \sigma^{a2}\right) \quad for\, t > 2 \tag{14}$$

$$\sigma^{a2} | k^a, \theta^a \; \sim \; InverseGamma(k^a, \theta^a) \tag{15}$$

$$p^a_{t,l} = S^a_t(l) - S^a_t(l-1) \tag{16}$$

$$S^a_t(l) = c(t,a) + (1 - c(t,a))S_U(l|t,a)$$

$$d^a_t(l) \sim Poisson\left(\lambda^a_t(l) \times n^a_t(l)\right) \tag{17}$$

$$\lambda^a_t(l) = 1 - \frac{S^a_t(l)}{S^a_t(l-1)} E^a_t(l) \tag{18}$$

All the distributions for the hyperparameters are here omitted to simplify notations. Equation (14) can be substituted with the first order autoregressive, as it will be shown in Section 6.

Equation (11) represents the first level of the hierarchical model while the second level is expressed both in (14) and (17). Let us stress, which will be better clarified in Section 3, that Equations (11) and (17) refer to different studies and data sets. M^a_t is obtained through the Regional Mortality Registry, while d^a_t is derived from a Local Cancer Registry (not necessarily in the same area, as will be clarified in Section 7). Furthermore, let us note that in Equation (17), d^a_t refers to deaths from any cause among people affected by the disease of interest, while in Equation (11), M^a_t refers to deaths from the cause of interest. Equation (18) represents a connection between the two studies, linking deaths from any cause to cause-specific survival S^a_t.

2.2. Age-period-cohort model for mortality

To project the cancer incidence, we first need to project the mortality counts. Following the model proposed by Bray [23], we specified a Bayesian APC model to project the future mortality counts. An APC model for M^a_t number of deaths at period t and age a is defined as follows:

$$M^a_t \sim \text{Poisson}\left(\mu^a_t\right) \tag{19}$$

$$\log\left(\mu^a_t\right) = \log\left(\text{Population}^a_t\right) + \alpha_{\text{age}=a} + \beta_{\text{period}=t} + \gamma_{\text{cohort}=n_a+t-a} \tag{20}$$

As is well known and as is easily found in [23], a cohort is linked to a period and age by the relation $Cohort = n_a + t - a$, where n_a is the number of age classes. We specify a Gaussian autoregressive prior model in the forward direction to smooth the effects on each timescale and to extrapolate the period and cohort effects. According to this model, each point (except for the first two points on each scale) is predicted by linear extrapolation from its two immediate predecessors in addition to a random error from a zero-mean Gaussian distribution in each of the three effects. In the autoregressive model, the first two parameters of age, period and cohort effects are given as noninformative priors (including a term for the hyperparameter to provide the correct likelihood). Supposing that projections are required for future periods, the prior distributions for the model parameters can be defined as follows.

$$\begin{aligned}
\alpha_j &\sim Normal\left(0, \frac{1}{\tau_\alpha}\right) & j = 1,2 \\
\alpha_a|\alpha_1, \dots ,_{a-1} &\sim Normal\left(2\alpha_{a-1} - \alpha_{a-2}, \frac{1}{\tau_\alpha}\right) & 3 \leqslant a \leqslant A \\
\beta_j &\sim Normal\left(0, \frac{1}{\tau_\beta}\right) & j = 1,2 \\
\beta_p|\beta_1, \dots ,_{p-1} &\sim Normal\left(2\beta_{p-1} - \beta_{p-2}, \frac{1}{\tau_\beta}\right) & 3 \leqslant p \leqslant P+N \\
\gamma_j &\sim Normal\left(0, \frac{1}{\tau_\gamma}\right) & j = 1,2 \\
\gamma_c|\gamma_1, \dots ,_{c-1} &\sim Normal\left(2\gamma_{c-1} - \gamma_{c-2}, \frac{1}{\tau_\gamma}\right) & 3 \leqslant c \leqslant C+N
\end{aligned}$$

where A is number of age classes, P is the number of periods and C is the number of cohorts. A prior distribution on precision parameter in the Gaussian distribution reflects prior beliefs concerning the smoothness of the parameters. We used noninformative hyperprior distributions so that the hyperparameters would be estimated solely from the data. Fitted and projected rates are then obtained through a recombination of the smoothed age, period and cohort effects, according to

$$\text{rate}_t^a = \exp\left(\alpha_{\text{age}=a} + \beta_{\text{period}=t} + \gamma_{\text{cohort}=A+t-a}\right) \qquad \forall a, t$$

The projected number of deaths are obtained from the knowledge of the population. Then, a back-calculation model can be used. To project the number of cancer cases, we run the APC component and then apply the back-calculation model at each iteration.

3. Data

All input data used in this study refer to the area of Florence and Prato districts, the two Italian provinces covered by the Tuscany Cancer Registry (TCR): 1.2 million residents, representing 33% of the entire Tuscany Region, which is located in Central Italy. We applied this model to the site of stomach cancer (ICD-IX code: 151) in women.

Cancer-specific mortality was extracted from the Regional Mortality Registry (RMR) by periods (19 calendar years, from 1987 to 2005) and age-classes (15–44, 45–54, 55–64, 65–74, 75–84 and 85–94). RMR collects all death cases that occurred among people residing in the Tuscany Region and among people who died in the Tuscany Region, even if they were not necessarily residents, following a national protocol.

Survival data for the same age-classes considered for cancer mortality in the TCR area were derived from patients diagnosed from 1987 to 2005 who were followed up until 31 December 2008. The TCR database provides the incident date for each patient, date and cause of death during the follow-up period or the number of patients alive at the end of follow-up. The survival data refer to all people who resided in the TCR area at the moment of incidence. TCR periodically updated the follow-up of the registered cases and checked patients' vital status to calculate their survival. Follow-up status was recorded using death certificates plus updating the last known date that the individual was known to be alive according to the civil registry.

Age-specific incident cancer cases in the TCR area observed from 1987 to 2005 were used to validate the incidence estimates. TCR periodically checks the quality and reliability of the incidence and survival data, according to the rules specified by the International Agency for Research on Cancer—IARC, Lion, France, which is followed up by the Italian Network of Cancer Registries—AIRTUM, Italy, with which the TCR is associated [25]. RMR checks the quality of the mortality data by comparing them with those published after years of delay by the National Institute of Statistics (ISTAT). These results are reported regularly in yearly reports [26].

4. Computational aspects

In Section 2, the three components of the specified model were illustrated, including the back-calculation equation, the survival distribution and the projections. The proposed model belongs to a class known as directed acyclic graphs (DAGs) as illustrated in Figure 4 and is as follows: rectangular nodes denote known constants; elliptical nodes represent either deterministic relationships (i.e. functions) or stochastic quantities, that is, quantities that require a distributional assumption; and stochastic dependence and functional dependence are denoted by single-edged arrows and double-edged arrows, respectively. Repetitive structures, such as loops, are represented by plates, which may be nested if the model is hierarchical [27]. DAGs can be used to pictorially describe a wide class of statistical models by describing the local relationships between quantities. DAGs communicate the essential structure of the model without requiring a large set of equations. The elegant underlying mathematical theory behind DAGs allows us to break down the analysis of these large and complex structures into a sequence of relatively simple computations. The proposed method can be implemented in the freely available software WinBUGS or OpenBUGS [27]. We implemented the algorithm using an interface of the CRAN R-packages [28]. We rewrote the hierarchical model in Just Another Gibbs Sampler (JAGS), a programme for Bayesian graphical modelling developed by Plummer in 2003 [29]. JAGS is written in C++, allowing for an object-oriented style that is extremely useful in this context. JAGS was originally created to be more similar to Classic Bugs while including improvements, and it has the advantages of running on all platforms and of interfacing with R. This programme is based on adaptive rejection Metropolis sampling [30] and is able to address nonlog-concave full conditional distributions. Adaptive rejection Metropolis samplers generalise adaptive rejection sampling to include a Hastings–Metropolis algorithm step to address nonlog-concave full conditional distributions.

After a sufficient number of burn-in iterations, the remaining samples from the Markov chain Monte Carlo (MCMC) simulations are used to obtain any function of the parameters of interest. To see how stable the final estimates are, multiple MCMC runs are conducted with different initial values and starting points. The convergence of the MCMC samples of the parameters after excluding the initial burn-in samples are monitored using the R package CODA [31].

To obtain the incidence estimates limited to the observed period of cancer mortality (up to 2005), we used the proposed model without the APC component because of the availability of the observed mortality data. Posterior estimations are obtained by running 60 000 iterations. The first 30 000 iterations were discarded as burn-in, and the results were obtained based on the remaining 30 000 iterations. To obtain incidence projections, mortality projections are needed. Therefore, we used the complete model, which includes the APC component, to project mortality. To take into account the uncertainty derived from mortality projections, we ran all of the components of the model together (APC component, survival component and back-calculation component). This application extended the necessary running time. The first 30 000 iterations were discarded as burn-in. The remaining 30 000 iterations were stored and used to obtain the posterior inferences.

5. Results

The Bayesian hierarchical model was fit using data stratified by six age-classes (15–44, 45–54, 55–64, 65–74, 75–84 and 85–94 years).

Table I. Observed and estimated relative survival at 5 years and 95% credible intervals.

Year	Observed	Estimated	95% CI	Age
1987–1990	0.42	0.37	0.25–0.49	15–44
1991–1995	0.25	0.38	0.27–0.49	15–44
1996–2000	0.48	0.40	0.29–0.52	15–44
2001–2005	0.56	0.41	0.28–0.54	15–44
1987–1990	0.32	0.34	0.26–0.43	45–54
1991–1995	0.39	0.35	0.29–0.43	45–54
1996–2000	0.36	0.36	0.29–0.44	45–54
2001–2005	0.40	0.37	0.28–0.46	45–54
1987–1990	0.35	0.37	0.32–0.43	55–64
1991–1995	0.39	0.39	0.34–0.44	55–64
1996–2000	0.46	0.40	0.35–0.46	55–64
2001–2005	0.41	0.42	0.35–0.49	55–64
1987–1990	0.33	0.32	0.28–0.36	65–74
1991–1995	0.33	0.33	0.29–0.36	65–74
1996–2000	0.32	0.33	0.29–0.37	65–74
2001–2005	0.37	0.34	0.29–0.39	65–74
1987–1990	0.26	0.25	0.22–0.29	75–84
1991–1995	0.24	0.26	0.23–0.29	75–84
1996–2000	0.25	0.27	0.23–0.30	75–84
2001–2005	0.31	0.28	0.24–0.32	75–84
1987–1990	0.10	0.14	0.10–0.19	85–94
1991–1995	0.14	0.15	0.11–0.20	85–94
1996–2000	0.15	0.15	0.10–0.20	85–94
2001–2005	0.11	0.15	0.10–0.19	85–94

Stomach cancer in women; Tuscany Cancer Registry data.

5.1. Survival curves

As a first step, we provide a mixture model estimation and analysis of stomach relative survival in women for the period from 1987 to 2005. Because of the necessity of 12 years of follow-up, survival parameters have been used to forecast the survival function in those cases for which a shorter period of observation was available (i.e. incident cases in the years from 2001 to 2005 for which a maximum of 8 years of follow-up were observable).

The observed and expected relative survival trends by follow-up periods for each age class confirm a good fit for the Bayesian cure model. Table I shows the observed and estimated 5-year relative survival with 95% credible intervals. The results show substantially stable relative survival in the first three age classes and an evident decrease in the older ones. All observed values belong to credible intervals, with the exception of two values in the first age class, probably because of the sparseness of the data.

Table II shows the temporal trends of a proportion of cured patients. According to each calendar year, the proportion of cured people increased in all age classes, except for the oldest class, which showed substantially stable results. The proportion of cured people appears to be similar in the youngest three age classes and decreases dramatically in the older ones. This age trend is consistent in all four of the calendar periods that were considered.

Table II. Proportion of cured people and 95% credible intervals.

Period	Age 15–44	Age 45–54	Age 55–64	Age 65–74	Age 75–84	Age 85–94
1987–1990	32.0 (21.6–45.0)	33.7 (25.7–42.9)	33.8 (25.8–39.4)	25.9 (19.5–30.4)	13.8 (5.7–21.0)	1.6 (0.1–7.9)
1991–1995	33.1 (23.1–45.1)	34.2 (27.3–41.8)	35.4 (28.1–39.8)	26.3 (20.9–30.3)	14.4 (6.2–21.4)	1.5 (0.1–7.7)
1996–2000	34.2 (23.5–46.6)	34.8 (27.6–42.6)	37.0 (28.9–41.8)	26.8 (21.4–31.4)	15.0 (6.4–22.4)	1.5 (0.05–7.6)
2001–2005	35.4 (23.0–49.8)	35.4 (26.4–44.9)	38.7 (28.8–45.3)	27.3 (21.1–33.8)	15.8 (6.4–24.1)	1.5 (0.05–7.7)

Stomach cancer in women; Tuscany Cancer Registry data.

Table III. Observed and estimated number of incident cases and 95% credible intervals.

Year	Observed	Estimated	95% CI	Age
1987–1990	4	5.51	2.29–9.42	15–44
1991–1995	5	5.12	2.06–9.02	15–44
1996–2000	5	3.35	0.94–6.99	15–44
2001–2005	3	2.43	0.15–6.87	15–44
1987–1990	13	12.97	8.40–18.27	45–54
1991–1995	12	13.67	9.18–18.78	45–54
1996–2000	13	11.83	7.56–16.51	45–54
2001–2005	8	6.47	2.19–12.98	45–54
1987–1990	32	33.05	25.67–41.41	55–64
1991–1995	30	30.47	24.09–37.20	55–64
1996–2000	25	23.86	17.70–30.55	55–64
2001–2005	19	17.17	8.79–26.92	55–64
1987–1990	74	68.18	58.21–78.74	65–74
1991–1995	64	62.89	54.10–71.99	65–74
1996–2000	52	51.7	43.41–60.50	65–74
2001–2005	45	39.06	28.01–51.43	65–74
1987–1990	114	122.96	110.93–136.79	75–84
1991–1995	102	102.47	91.19–114.61	75–84
1996–2000	85	84.38	74.33–95.35	75–84
2001–2005	74	74.27	60.65–88.91	75–84
1987–1990	50	42.04	32.37–51.91	85–94
1991–1995	60	53.1	44.04–62.07	85–94
1996–2000	56	53.97	45.75–61.95	85–94
2001–2005	45	41.62	29.13–53.82	85–94

Stomach cancer in women; Tuscany Cancer Registry data.

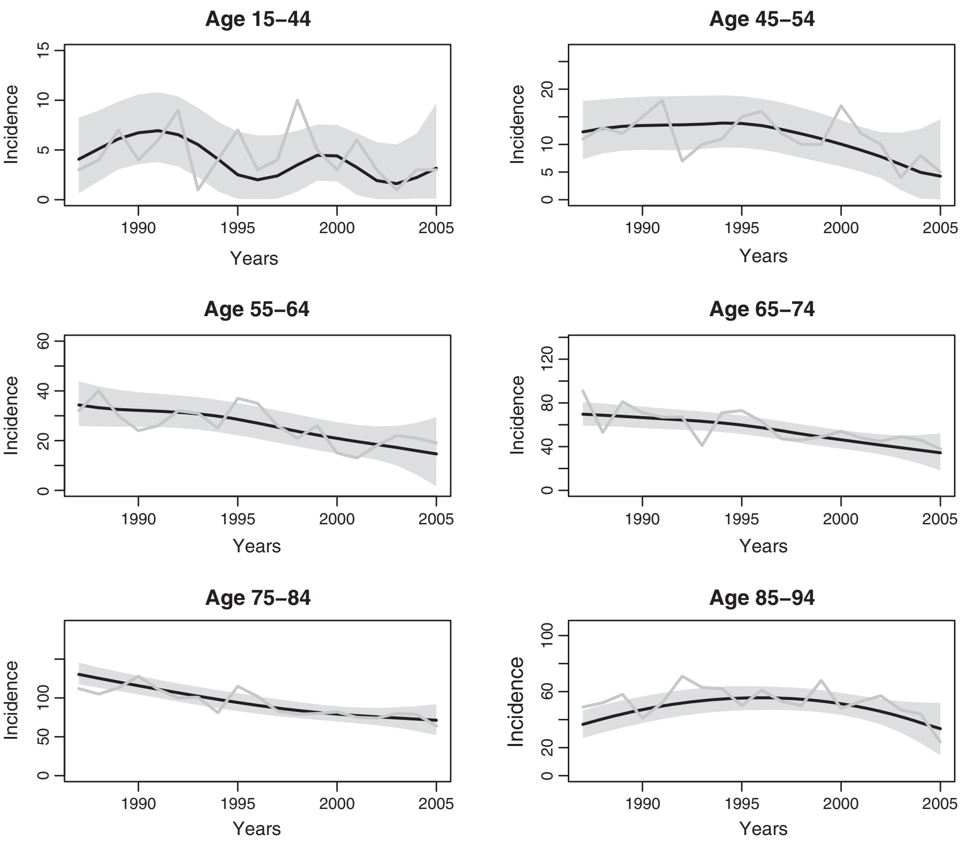

Figure 1. Observed (grey line) and estimated (black line) time trend incident cases; 95% credible interval (grey band). Tuscany Cancer Registry data.

5.2. Incidence

In Table III, the observed and estimated numbers of new cancer cases are reported with 95% credibility intervals. The data are presented as the means over 5-year period for each age class (apart from the first period from 1987 to 1990). The estimated incidence cases were in good agreement with the observed data. The 95% credible intervals contained the observed data in all age classes and calendar years analysed.

Figure 1 reports the observed and estimated number of incident cases for each year by age class, and 95% credible intervals are reported. The fit seems to be quite good for all age classes. Although the y scale is specific for each age class to allow for a good presentation of the age-specific results, it is easy to note that, proportionally, the credible intervals were larger for the younger ages. Wider credible intervals in the younger age classes may be due to the small number of cases and may adequately indicate the real uncertainty of the data, as confirmed by the variability of the observed cases. All age classes showed a decreasing trend.

5.3. Forecast and validation

To forecast cancer incidence beyond the years of the observed mortality and survival, we used the APC and cure model, respectively. The fit of the two models was quite good. To validate the mortality projections, we applied the APC model to a restricted data set of observed data (until 2000).

Figure 2 shows the validation of mortality projections. Estimates and credible intervals for the period from 1987 to 2000 are not shown because the observed mortality data are used in the back-calculation model. To validate the mortality projections, we compared the projected and observed mortality relative to the period from 2001 to 2005, as obtained through the APC model applied to the mortality counts relative to those from 1987 to 2000. The projected figures are very close to the observed ones. Credible intervals grow rapidly as the projected period grows, highlighting the large uncertainty of long-term period projections.

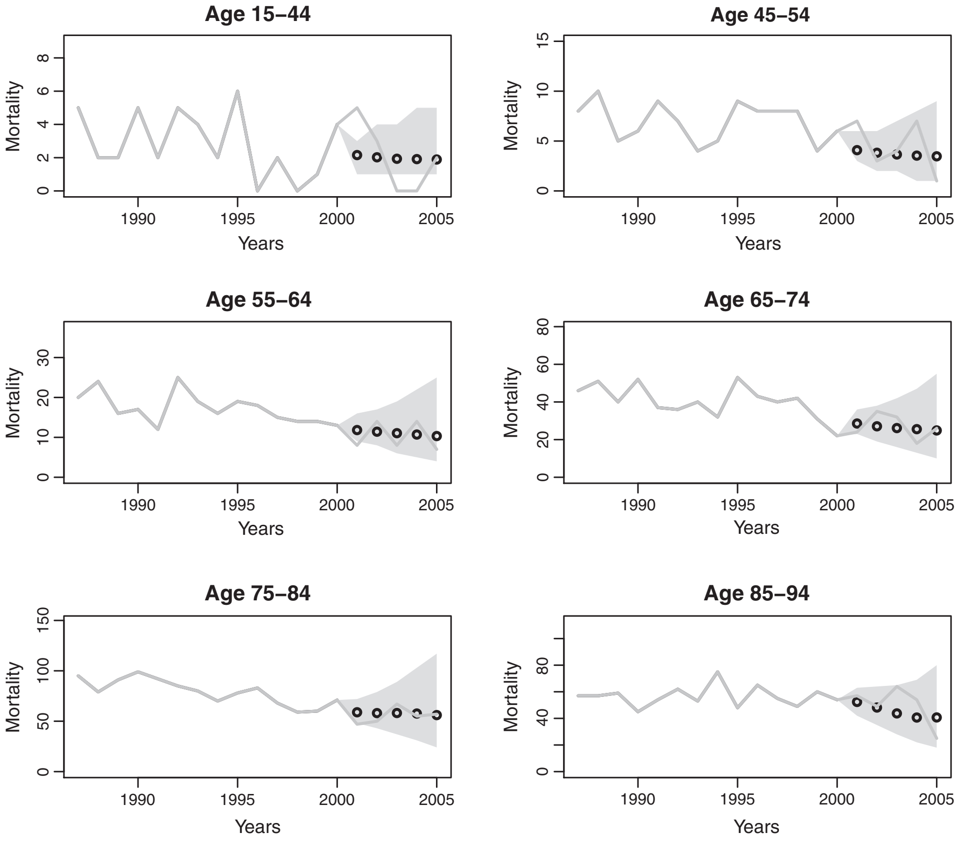

Figure 2. The grey line represents the observed time trend mortality from 1987 to 2005. The black points represent the projected mortality for the period 2001–2005 obtained through the application of the APC model based upon data up to 2000. The grey band represents the 95% credible intervals. The data are referred to the Tuscany Mortality Registry.

To validate the incidence projections, we followed a similar procedure as that for mortality in which we restricted the input data (mortality and survival) to the period from 1987 to 2000. Figure 3 shows the estimated and projected trends, together with the observed incidence, for each age class. Credible intervals tended to increase when projecting because of the capacity of the model to take into account of the uncertainty of all the components of the hierarchical model.

5.4. *Convergence*

Convergence of the estimates has been checked by the test of convergence diagnostics proposed by Geweke [32] and implemented in the software CODA [31].

The z-score for the test of the equality of the means results in no statistical significance for more than 95% of the tested parameters in the back-calculation model and for more than 92% of the tested parameters (about double parameters) in the back-calculation joint with APC model.

6. Sensitivity analysis

To measure the robustness of the results regarding the selection of prior distributions, a sensitivity analysis with regard to the choice of prior distributions was carried out. Sensitivity to assumptions needs to be investigated, both on the specification of the prior distribution and on the specification of hyperparameters.

The Poisson distribution is the most adequate distribution to describe incidence counts, especially when a small number of cases are observed, as occurred in our application of the first age class. Because the incidence counts were within a convolution equation dependent upon mortality counts and because the disease was described by a Poisson distribution for mortality counts in the first level of the hierarchical

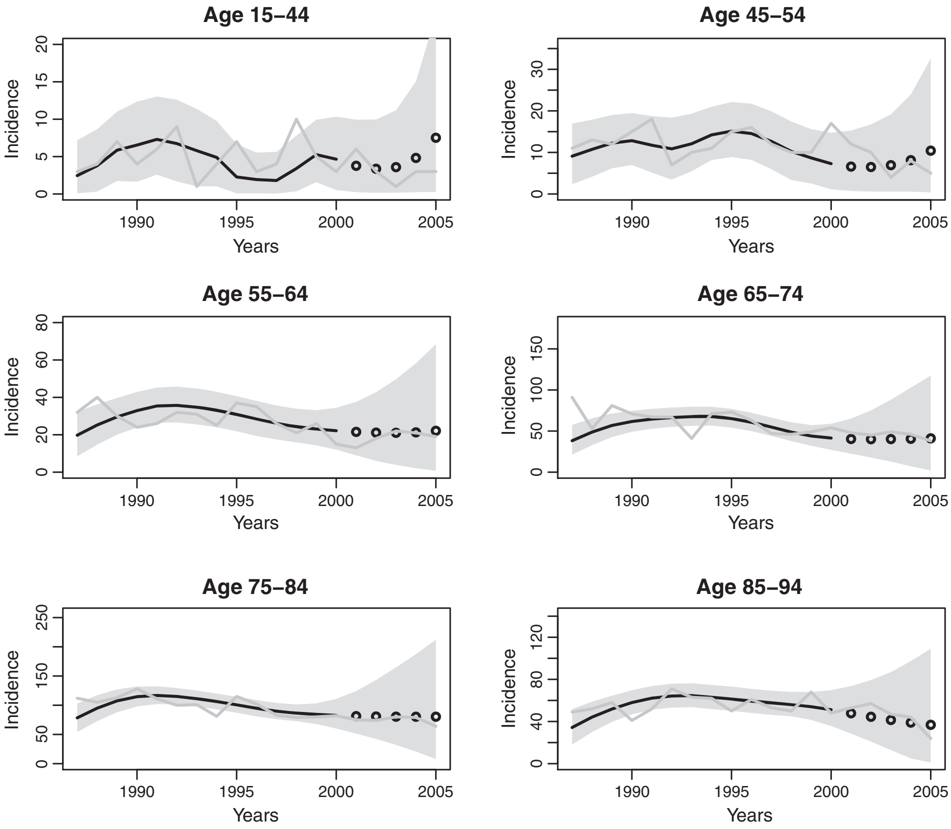

Figure 3. The grey line represents the observed time trend incident cases from 1987 to 2005. The black line represents the incidence estimates for the period 1987–2000. The black points represent the projected incidence for the period 2001-2005. Both estimates and projections are obtained through the application of the back-calculation model based upon data up to 2000. The grey band represents the 95% credible intervals. The data are referred to the Tuscany Cancer Registry.

model, we chose to give priority to the autoregressive structure on the second hierarchical level, easily described by a Gaussian distribution.

We evaluated the autoregressive structure for this incidence by centering the Gaussian distribution of the incidence on the previous year alone, instead of on the two immediate predecessors. This choice did not significantly change the final estimates; for all age classes except the first, the posterior estimates differed in few cases. In the first class, the estimated trend was less smoothed, and the five observed values were outside of the credible intervals.

Once the prior distribution in (14) was selected, both the shape and inverse scale parameters of the Gamma distribution were considered to vary over the range [0.001; 0.1]; the posterior estimates changed by less than one unit in all age classes. The precision of the Gaussian distribution in the autoregressive structure did not influence the final estimates, and the posterior estimates for precision remain fairly stable.

In past studies, sensitivity analysis has revealed that the final estimates are sensitive to the choice of survival distribution [5, 6, 11]; this constituted the primary drawback to the back-calculation models previously proposed. A possible solution to this flaw is the assumption of a more complex and flexible model. Several distributions that are more frequently used for survival data were investigated, and the choice of a Bayesian mixture cure model became the best choice in terms of comparing observed and estimated values, both for survival and incidence estimates. By assuming a survival model with three components, we are able to allow the data to suggest the best structure so that the model may be applied to other cancer sites, reducing the risk of misspecification. The final model results in a type of model averaging that is able to account for a source of uncertainty that an analysis based on model selection would ignore.

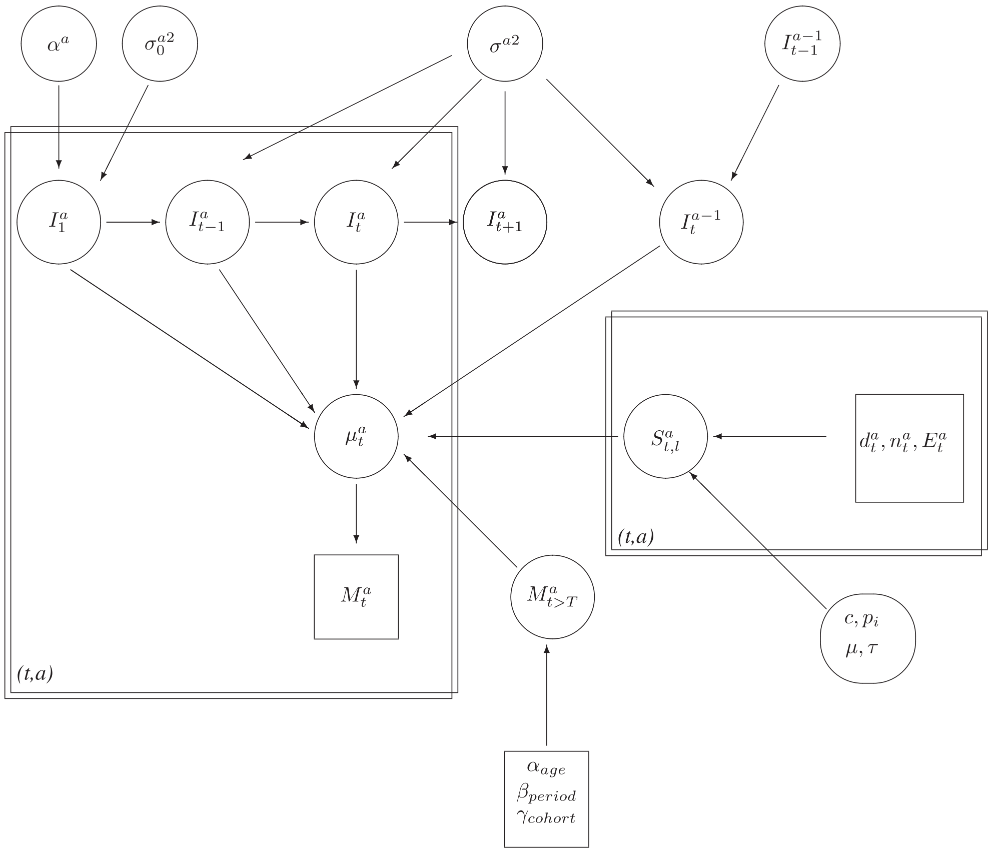

Figure 4. Directed acyclic graph for back-calculation model.

We ran the model by choosing only one of the three components at a time. Both the evaluation of the fit and the deviance information criterion (DIC) confirmed that the three-component mixture model was the best choice with the lowest DIC value.

Different values for precision in the Gaussian hyperprior distribution in Equations (12)–(14) have been tested, including 10, 100 and 1000; both the shape and inverse scale parameters in the Gamma hyperprior distribution were considered to vary over the range [0.001; 0.01]. The final estimates were not sensitive to changes in the hyperparameters. Regarding APC model, final estimates resulted quite stable when we consider small changes for values for the mean of the normal distributions for the first two parameters for age, period and cohort and for the shape and inverse scale parameters of the Gamma distributions.

The complexity of the proposed model requires substantial computational effort. A possible solution could be fitting the model by running the survival model and the APC separately. The model structure would not change because the two components do not share information as they are based on different studies. As a third step, back-calculation model could be applied by using survival distribution and mortality counts obtained in the first two steps. We consider the possibility of defining our problem an advantage based upon a unique DAG model, as illustrated in Figure 4, because our model allows for the correct quantification of these uncertainties. Moreover, the advantage of DAG models is their possibility of reducing complexity.

7. Discussion

The aim of this paper is to show how the relationships between incidence, survival and mortality can be used to estimate the incidence cases in areas for which they were not recorded. Moreover, incidence can be forecasted when mortality has not yet been observed. Sometimes, the incidence and survival data are obtained from a register in a selected population (generally defined according to the area of residence). If the population covered by the register is representative of the general population for available health care facilities (both diagnostic and therapeutic), then the observed survival data may be reasonably extended to the latter. Such an extension is not generally valid for incidence, which is strongly affected by the geographical variability of risk factors related to both environment and lifestyle. The survival data can be used jointly with the observed mortality rates to yield incidence estimates in the population as a whole.

One possible application of our model is to estimate cancer incidence in areas not covered by cancer registration, such as the case of the Tuscany Region. As cancer survival data are not yet available in the Tuscany Region, the Florence-Prato Cancer Register can be chosen as an alternative source of data, and corresponding survival probabilities can be assumed for cancers occurring in the Tuscany Region. The survival data from the Florence-Prato district can be used jointly with mortality counts from the Tuscany Region to yield regional incidence estimates.

Applying our model to women with stomach cancer, the ability to correctly reconstruct the age-specific incidence trend starting from mortality and survival data is demonstrated. Moreover, methods for taking into account all of the uncertainties in the modelled input data are discussed. It is of fundamental importance to consider the uncertainty around the point estimates when they are used for planning public health interventions. This becomes particularly important when projections are concerned, because of the fast rate of enlarging of the credible intervals.

The reliability of the estimates obtained by applying the proposed model is strictly dependent on the quality of the data. Ascertaining the quality and completeness of the data with regard to possible sources of error and distortion should be a preliminary step in using the method proposed. TCR periodically checks the quality and reliability of the survival and incidence data, according to international rules [25]. Use of an excessively optimistic survival curve would lead to an overestimation of the incidence rates. In the comparison of estimated and observed cancer cases, any possible under-reporting of new cases must also be taken into account. Incompleteness of the registration of incidence would lead to a corresponding underestimation of the incidence. The completeness of registration can be evaluated by the means of the percentage of cases detected through the death certificate (death registration only cases, DCO). The lower this value, the better the quality of register incidence data. In our case, the percentage of the DCO for the stomach cancer site in women is approximately 2%, similar to that of the nationally reported data [2]. As far as mortality data are concerned, RMR checks the quality of the mortality data through comparison with data published by ISTAT after several years of delay.

The novelty of the proposed model is its ability to take into account the uncertainty of the survival estimates that are required for the computation of incidence estimates in areas not covered by the cancer registry and of the survival and mortality projections required for the forecasting of incidence. The 95% credible intervals are presented in order to quantify this variability. The quantification of the variability is needed mostly if the number of future new cancer cases are used by the policy maker in planning health system interventions.

In traditional back-calculation methods, such as the MIAMOD model [5, 6], the uncertainty within the incidence estimates is usually not presented. This may be due to the inability of such models to take into account all of the uncertainty from the modelled input data (mortality and survival). The bootstrap method has been proposed to provide confidence intervals for back-calculation estimates [33]. The lack of information about uncertainty becomes crucial when projections are concerned and when estimates are based on a low number of deaths.

Like many Bayesian models, our proposal appears to be very heavily parameterized; more parsimonious models have been investigated. The proposed method provides better posterior estimations with respect to the restricted model as in Mezzetti and Robertson [11]. Moreover, a critical point in back-calculation methods is given by the survival function. The survival distribution has to be properly modelled to smooth and project the observed data; restricted models fail to consider the relative uncertainty. The proposed method for survival was developed specifically for relative survival. The relative survival is the preferred measure of cancer survival when we analyse population-based data when the cause of death is not completely reliable. We chose to model the survival function by using a Bayesian-cured survival model [20]. Our proposed model is a three-component mixture of the three most commonly used distributions in survival analysis: log-normal, log-logistic and Weibull. The mixture of different components is used to increase the flexibility of accommodating distributions with different shapes and tails. The idea behind the proposed mixture cure models is to start with a complex model and to allow the data to suggest the weight of different distributions and the determination of an optimal combination of components. The mixture cure model can also be thought of as a special case of model averaging. When faced with several candidate models, the analyst can either choose one model or average several models, as the posterior estimates suggest. For example, other applications of this concept we performed (such as colorectal cancer) demonstrated a greater weight for the log-normal and log-logistic distributions, with a low weight for the Weibull distribution.

Because of the aging of the population, it is interesting to consider careful estimates for older ages. We found that by using an open upper age class (85+), the incidence estimates were severely underestimated. This is most likely due to the progressive and dramatic decrease of the relative survival at late ages in life. We propose to solve this problem by capping the upper age class at 94. We did not model further age classes because of sparse data.

To project cancer incidence into the future, we previously need to project mortality counts from which the back-calculation model starts. We applied the Bayesian APC model proposed by Bray *et al.* [23] with autoregressive smoothing to make mortality projections. The proposed Bayesian approach attributes separate effects to age, period and cohort, and extrapolates these effects to make projections. Hakulinen and Dyba [34] argued against the use of APC models for making projections on the grounds that the considerable increase in the number of parameters over simpler linear models results in large standard errors and decreased precision. However, as Berzuini and Clayton [35] noted, 'the a priori belief in smoothness that is inherent in the Bayesian approach leads to more precise estimates of rates than the corresponding maximum likelihood estimates'. In the Bayesian version of this method, the most appropriate degree of smoothing can be determined from the data. The Bayesian model is also more flexible than the linear power models [34, 36] because it copes with both increasing and decreasing trends. It is important to note that the model hinges on the assumption that past trends will continue into the future. This may be a strong assumption when long-term projections are computed. In our case, we projected into the future for a limited number of years, obtaining short-term cancer projections. In this case, this assumption seems to be reasonable.

The model we proposed gave good results, with the estimates being very close to the observed data contained in the 95% credible intervals for all age groups, considering 5-year periods, which are usually reported in several studies. We compared our estimates with those obtained by the traditional back-calculation methods (MIAMOD), which are available online (www.tumori.net). The comparison is not the most formal one, because the MIAMOD estimates were not produced using the same input data but were retrieved on the web. We applied the estimated MIAMOD incident rates to our population in order to obtain the number of incident cases for comparison. The comparison showed that the MIAMOD estimates seem to be more smoothed during the observed period with a slight underestimation, because of the parabolic shape of the observed data in those years. The projected number of cases showed very similar results between the two methods.

We believe the method proposed will have a wide application in all geographical areas, even where the incidence is collected only for few areas and for a particular disease. Moreover, in many countries, including Italy, the emerging necessity for the quantification of cancer costs and the consequent resource allocation demands a method that is able to estimate cancer incidence while correctly quantifying uncertainties.

Acknowledgements

The authors would like to thank Guido Miccinesi, Chris Robertson, Binbing Yu, Silvia Francisci, Roberta De Angelis, Riccardo Capocaccia and two unknown referees for the helpful comments.

References

1. Brown ML, Lipscomb J, Snyder C. The burden of illness of cancer: economic cost and quality of life. *Annnual Review of Public Health* 2001; **22**:91–113. DOI: 10.1146/annurev.publhealth.22.1.91
2. AIRTUM Working Group. Italian cancer figures, report 2011: survival of cancer patients in Italy. *Epidemiologia e Prevenzione* 2011; **35**(5-6(3)):1–200.
3. Fulton JP, Howe HL. Evaluating the incidence-mortality ratios in estimating completeness in cancer registration. In *Cancer Incidence in North America, 1988-1991*, Howe HL (ed.) North American Association of Central Cancer Registries: Sacramento, CA, 1995; VI 1–VI 9.
4. Colonna M, Grosclaude P, Faivre J, Revzani A, Arveux P, Chaplain G, Tretarre B, Launoy G, Lesec'h JM, Raverdy N, Schaffer P, Buémi A, Ménégoz F, Black RJ. Cancer registry data based estimation of regional cancer incidence: application to breast and colorectal cancer in French administrative regions. *Journal of Epidemiol and Community Health* 1999; **53**(9):558–564.
5. Verdecchia A, Capocaccia R, Egidi V, Golini A. A method for the estimation of chronic disease morbidity and trends from mortality data. *Statistics in Medicine* 1989; **8**(2):201–216. DOI: 10.1002/sim.4780080207
6. De Angelis G, De Angelis R, Frova L, Verdecchia A. MIAMOD: a computer package to estimate chronic disease morbidity using mortality and survival data. *Computer Methods and Programs in Biomedicine* 1994; **44**(2):99–107.

7. Sant M, Francisci S, Capocaccia R, Verdecchia A, Allemani C, Berrino F. Time trends of breast cancer survival in Europe in relation to incidence and mortality. *International Journal of Cancer* 2006; **119**(10):2417–2422. DOI: 10.1002/ijc.22160
8. Mariotto AB, Yabroff KR, Feuer EJ, De Angelis R, Brown M. Projecting the number of patients with colorectal carcinoma by phases of care in the US: 2000-2020. *Cancer Causes and Control* 2006; **17**(10):1215–1226. DOI: 10.1007/s10552-006-0072-0
9. Capocaccia R, Buzzoni C, Grande E, Inghelmann R, Bellù F, Cassetti T, de Dottori M, Donato A, De Lisi V, Falcini F, Federico M, Ferretti S, Fusco M, Giacomin A, Guzzinati S, Mangone L, Piffer S, Rosso S, Sechi O, Tagliabue G, Tumino R, Vercelli M, Vitarelli S. Estimated and observed cancer incidence in Italy: a validation study. *Tumori* 2007; **93**(4): 387–391. DOI: 10.1700/294.3482
10. Ventura L, Miccinesi G, Sacchettini C, De Angelis R. Estimating all cancers incidence with the MIAMOD model: a new method to include multiple tumors. *Tumori* 2012; **98**(3):296–302. DOI: 10.1700/1125.12395
11. Mezzetti M, Robertson C. A hierarchical Bayesian approach to age-specific back-calculation of cancer incidence rates. *Statistics in Medicine* 1999; **18**(8):919–933. DOI: 10.1002/(SICI)1097-0258(19990430)18:8<919::AID-SIM89>3.0.CO;2-7
12. Kleinbaum DG, Klein M. *Survival Analysis—A Self Learning Text, Third Edition*. Springer: New York, 2011.
13. Boag JW. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)* 1949; **11**(1):15–53. DOI: 10.2307/2983694
14. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* 1952; **47**(259):501–515. DOI: 10.1080/01621459.1952.10501187
15. De Angelis R, Capocaccia R, Hakulinen T, Soderman B, Verdecchia A. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine* 1999; **18**(4):441–454. DOI: 10.1002/(SICI)1097-0258(19990228)18:4<441::AID-SIM23>3.0.CO;2-M
16. Seppä K, Hakulinen T, Kim HJ, Läärä E. Cure fraction model with random effects for regional variation in cancer survival. *Statistics in Medicine* 2010; **29**(27):2781–2793. DOI: 10.1002/sim.4046
17. Peng Y, Dear KBG. A nonparametric mixture model for cure rate estimation. *Biometrics* 2000; **56**(1):237–243. DOI: 10.1111/j.0006-341X.2000.00237.x
18. Sy JP, Taylor JMG. Estimation in a Cox proportional hazards cure model. *Biometrics* 2000; **56**(1):227–236. DOI: 10.1111/j.0006-341X.2000.00227.x
19. Lu W, Ying Z. On semiparametric transformation cure models. *Biometrika* 2004; **91**(2):331–343. DOI: 10.1093/biomet/91.2.331
20. Yu B, Tiwari RC. A Bayesian approach to mixture cure models with spatial frailties for population-based cancer relative survival data. *Canadian Journal of Statistics* 2012; **40**(1):40–54. DOI: 10.1002/cjs.10135
21. Clayton D, Schifflers E. Models for temporal variation in cancer rates. II: age-period-cohort models. *Statistics in Medicine* 1987; **6**(4):469–481.
22. Bashir SA, Estève J. Projecting cancer incidence and mortality using Bayesian age-period-cohort models. *Journal of Epidemiology and Biostatistics* 2001; **6**(3):287–296. DOI: 10.1080/135952201317080698
23. Bray I. Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 2002; **51**(2):151–164. DOI: 10.1111/1467-9876.00260
24. Begg CB, Schrag D. Attribution of deaths following cancer treatment. *Journal of the National Cancer Institute* 2002; **94**(14):1044–1045. DOI: 10.1093/jnci/94.14.1044
25. Curado MP, Edwards B, Shin HR, Storm H, Ferlay J, Heanue M, Boyle P (eds). *Cancer Incidence in Five Continents, Vol. IX*. IARC Scientific Publication: Lion, 2007.
26. Chellini E, Giovannetti L, Sorso B, Fornai M G, Martini A, Querci A, Seniori Costantini A (eds). *Morti per causa anno 2006*. Edizioni Regione Toscana: Firenze, 2007.
27. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: evolution, critique and future directions. *Statistics in Medicine* 2009; **28**(25):3049–3067. DOI: 10.1002/sim.3680
28. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria, 2012. http://www.R-project.org/,ISBN3-900051-07-0 [Accessed on 15 June 2013].
29. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Hornik K, Leisch F, Zeileis A (eds). Production: Achim Zeileis: Vienna, Austria, 2003; 1–10.
30. Gilks WR, Best NG. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1995; **44**(4):455–472.
31. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News* 2006; **6**(1):7–11.
32. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculating posterior moments. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, Bernardo JM, Berger J, Dawid AP, Smith JFM (eds). Clarendon Press: Oxford, 1992; 169–193.
33. Gigli A, Verdecchia A. Uncertainty of AIDS incubation time and its effects on back-calculation estimates. *Statistics in medicine* 2000; **19**(2):175–189. DOI: 10.1002/(SICI)1097-0258(20000130)19:2<175::AID-SIM299>3.0.CO;2-S
34. Hakulinen T, Dyba T. Precision of incidence predictions based on Poisson distributed observations. *Statistics in Medicine* 1994; **13**(15):1513–1523. DOI: 10.1002/sim.4780131503
35. Berzuini C, Clayton D. Bayesian analysis of survival on multiple time scales. *Statistics in Medicine* 1994; **13**(8):823–838. DOI: 10.1002/sim.4780130804
36. Dyba T, Hakulinen T, Päivärinta L. A simple non-linear model in incidence prediction. *Statistics in Medicine* 1997; **16**(20):2297–2309. DOI: 10.1002/(SICI)1097-0258(19971030)16:20<2297::AID-SIM668>3.0.CO;2-F