Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting annual lung and bronchus cancer deaths using individual survival times

Duk Bin Jun[a,*], Kyunghoon Kim[a], Myoung Hwan Park[b]

[a] *KAIST College of Business, Seoul, Republic of Korea*
[b] *Hansung University, Seoul, Republic of Korea*

ARTICLE INFO

Keywords:
Health forecasting
Cancer
Survival analysis
Weibull mixture model
Unobserved heterogeneity

ABSTRACT

Accurate forecasts of the numbers of cancer deaths are critical not only for allocating government health and welfare budgets, but also for providing guidance to the related industries. We suggest a framework for predicting the annual numbers of cancer deaths by modeling individual survival times. A Weibull mixture model with individual covariates and unobserved heterogeneity is proposed for examining the effects of demographic variables on individual survival times and predicting the annual number of cancer deaths by adopting a bottom-up strategy. We apply the suggested framework to a survival analysis of lung and bronchus cancer patients in the United States and provide a comparison with the forecast results obtained from previous studies. A comparison of our results with those of various benchmarks shows that our proposed model performs better for predicting annual numbers of cancer deaths. Furthermore, by segmenting patients based on age, sex, and race, we are able to specify the differences between groups and assess the group-specific survival probabilities within a given period. Our results show that older, female, and white patients survive significantly longer than younger, male, and black patients. Also, patients diagnosed in recent years survive significantly longer than those diagnosed a long time ago.

1. Introduction

Worldwide, approximately one death in eight is due to cancer. The estimated total number of cancer deaths worldwide in 2008 was 7.6 million, and the growth and aging of the population mean that this number is expected to reach 13.2 million in 2030 (American Cancer Society, 2011). It is predicted that around 585,720 people will die of cancer in the United States in 2014, a rate of about 1,600 per day (Siegel, Jiemin, Zou, & Jemal, 2014). Since billions of dollars are spent on research, treatment, prevention, and other cancer-related costs, accurate predictions of the numbers of cancer deaths are important for effective planning, resource allocation, and communication (Tiwari et al., 2004). Accurate predictions of the numbers of cancer deaths are beneficial for the public sector because they enable more precise allocations of health and welfare budgets. If the predicted number of deaths does not match the actual number of cancer deaths, then the planned budgets could be either too high or inefficient, leading to difficulties for governments. Accurate predictions are also critical in the private sector. For example, accurate forecasts of the numbers of cancer deaths allow insurance firms to predict the need for cancer insurance.

For these reasons, there has been a considerable amount of research in recent years relating to the forecasting of the numbers of cancer deaths using aggregate-level data. Chen et al. (2012) compared the levels of accuracy

[*] Corresponding author.
E-mail address: dbjun@business.kaist.ac.kr (D.B. Jun).

http://dx.doi.org/10.1016/j.ijforecast.2015.05.002

of five time series models for four-year-ahead projections of the numbers of cancer deaths. The models were as follows: (1) the state space model, (2) the Bayesian state space model (Schmidt & Pereira, 2011), (3) two versions of the joinpoint regression model (Kim, Fay, Feuer, & Midthune, 2000), (4) the Nordpred model (Møller et al., 2003), and (5) a vector autoregressive model using the Hilbert–Huang transform (Huang et al., 1998). Chen et al. (2012) showed that the joinpoint model with the modified Bayesian information criterion had the smallest error among the various models tested. Tiwari et al. (2004) improved the existing state space model and revealed that the average squared deviations across cancer sites for the new model were substantially lower than those of other benchmark models. However, so far, such models have used only aggregate-level mortality data. On the other hand, Verdecchia, De Angelis, and Capocaccia (2002) proposed the prevalence, incidence, and analysis model (PIAMOD) for predicting numbers of deaths using incidence data. The model first fitted the incidence based on age, period, and cohort, then derived the numbers of deaths.

Some studies have used survival analysis to focus on the individual hazard or survival rate. The most prominent model in survival analysis is the proportional hazards regression model, proposed by Cox (1972). Since then, the Cox model has been used widely for specifying a linear relationship between hazard or survival rates and covariates in a variety of fields, such as engineering, economics, and sociology. In biometrics, there have been studies of methods for modeling an individual's lifetime (Hakulinen & Tenkanen, 1987; Prentice & Kalbfleisch, 1979), which have generally involved the use of parametric models. One of the most popular of these is the Weibull model (Cox, 1972). Royston and Parmar (2002) developed flexible parametric models based initially on the assumption of a proportional hazards scaling of covariate effects. This class of models was based on a transformation of the survival function using a link function.

When subjects are taken from a heterogeneous population, however, a model that allows us to deal with the unobserved heterogeneity should be considered. If we ignore heterogeneity, our estimates could be spurious (Lancaster, 1992). Lancaster (1979) used a Weibull-gamma mixture or finite mixture model using Bayesian methods to capture the unobserved heterogeneity. Frailty models have also been used widely to address the issue of heterogeneity. The models assume that different individuals have different frailties; frailer individuals tend to die earlier than those who are less frail. Vaupel, Manton, and Stallard (1979) were the first to introduce the concept of frailty and apply it to population data. When the population is a mixture of susceptible and non-susceptible individuals, the frailty model can be used to extend the cure model, leading to the so-called cure frailty model. Duchateau and Janssen (2007) introduced several examples of frailty models, from those assuming parametric distributions, including gamma, Weibull, and lognormal distributions, to nonparametric frailty models. They also reviewed frailty models that remove the assumption that the frailty is constant over time, the so-called time-varying frailty models. These hazard or frailty models can explain individual survival patterns. However, there has been little or no research into the use of individual survival times for predicting total numbers of failures or deaths.

In this paper, we propose a Weibull mixture model with individual covariates and unobserved heterogeneity for examining the effects of demographic variables on individual survival times and predicting the annual numbers of cancer deaths by adopting a bottom-up strategy. Our model is applied to data on lung and bronchus cancer patients in the United States, and specifies the relationship between survival times, demographic variables (age, sex, race, and registries), and incidence-related variables (year of diagnosis and stage of tumor progression). We then obtain four-year-ahead forecasts for 2006, 2007, 2008, 2009, 2010, and 2011. The performance of our model is compared to those of other benchmarks that previous research has shown to be accurate in predicting the numbers of deaths. Moreover, we segment the whole patient dataset into several groups based on age, sex, and race, in order to examine the heterogeneity between groups and calculate group survival probabilities for the next three years.

The remainder of the paper is organized as follows. In Section 2, we provide a description of the Surveillance, Epidemiology, and End Results (SEER) dataset used in the analysis. In Section 3, we develop our modeling framework and explain how the annual number of deaths is forecast. In Section 4, we describe our empirical analysis and specify the parameter estimates and forecast results. This is followed by a discussion of the implications of our study, and we conclude with a summary of this research and its limitations, as well as a review of its contributions to the public and private sectors.

2. Data description

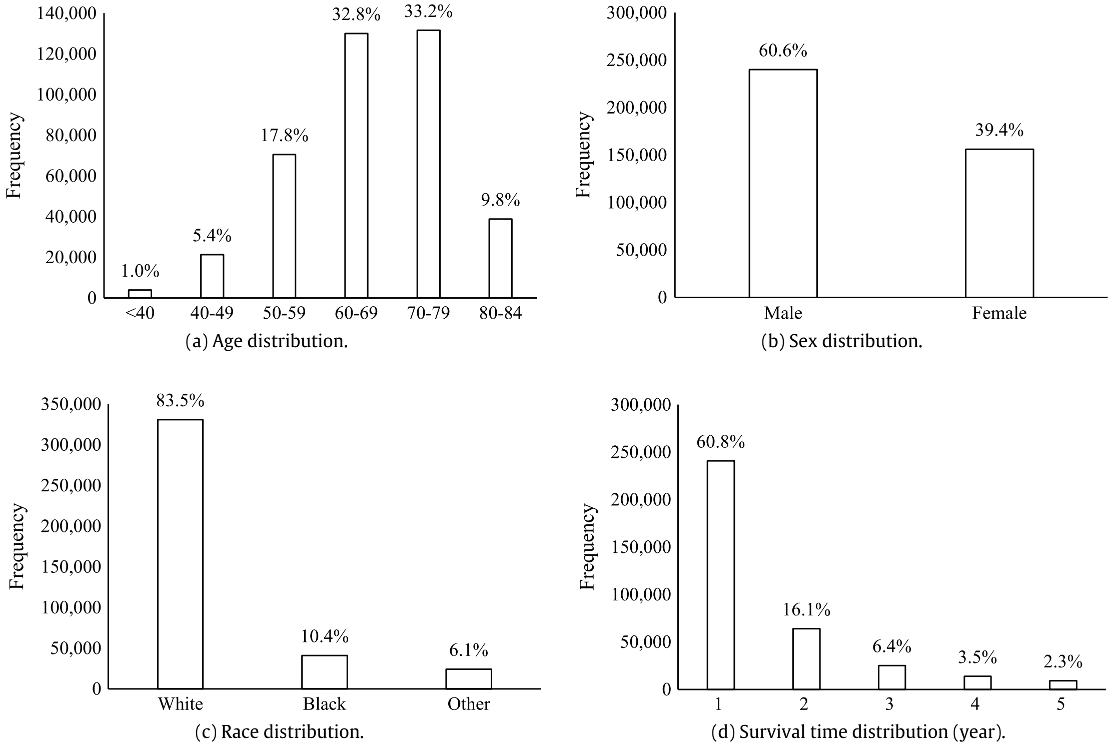
The dataset is retrieved from the SEER database of the National Cancer Institute.[1] It consists of information about patients who were diagnosed with cancer between 1973 and 2011 from nine registries[2] in the United States, representing approximately 10% of the US population. We make use of the data only until 2011 because the incidence and mortality data are generally available with a lag of three to four years, due to the time required for data collection, compilation, quality control, and dissemination (Siegel et al., 2014). The data gathered range from demographic variables, such as age at diagnosis, sex, race, marital status, and registry, to cancer incidence-related variables, such as year of diagnosis, survival time, and stage of tumor progression. The data are right-censored, with the follow-up cutoff date fixed at December 31, 2011.

We focus on patients diagnosed with lung and bronchus cancers.[3] There are a total of 396,202 patients; however,

[1] http://www.seer.cancer.gov/. Released April 2014, based on the November 2013 submission.

[2] The registries are from Atlanta, Connecticut, Detroit, Hawaii, Iowa, New Mexico, San Francisco-Oakland, Seattle-Puget Sound, and Utah.

[3] These cancer types are classified as "Lung" in the "CS schema v0204" classification. This classification includes the types of cancer defined by the ICD-10 codes C340, C341, C342, C343, C348, C349, and D022.



(a) Age distribution.

(b) Sex distribution.

(c) Race distribution.

(d) Survival time distribution (year).

Fig. 1. Age, sex, race, and survival time distribution of lung and bronchus cancer patients.

those registered in the Seattle-Puget Sound and Atlanta areas are excluded from our analysis because their data were not gathered from the origin year of 1973.[4] Also, we do not include patients over the age of 85 because their deaths may not be cancer-related. When patients were registered, most of their tumors (99%) were found to be malignant. As Fig. 1(a) shows, 80% of the patients were between 50 and 80 years old. Fig. 1(b) and (c) show that the number of males was one and a half times the number of females, and the proportion of white patients was 84% of the total. In Fig. 1(d), it can be seen that 89% of patients did not survive for more than five years, while 61% died within one year, and 83% died within three years.

3. Model

3.1. Specification

In survival analysis, many studies have modeled survival times using an exponential distribution. Since the exponential model defines the survival probability based on a single parameter, it has the advantage of simplicity. However, it has the limitation that survival times sampled from the exponential distribution have a constant hazard rate, which assumes that all subjects are equally likely to die, regardless of how many years they have survived. This is undoubtedly unrealistic in many situations. On the other hand, the Weibull model can account for hazard rates both increasing and decreasing. The Weibull distribution is shaped by two parameters: the rate parameter (denoted by λ, $\lambda \geq 0$), which is the inverse of the scale parameter; and the shape parameter (denoted by c, $c \geq 0$), which provides flexibility in the hazard rate. The probability density function (PDF) and the survival probability of the Weibull distribution with λ and c are given by:

$$f\left(t|\lambda, c\right) = \lambda c t^{c-1} \exp\left(-\lambda t^{c}\right) \tag{1}$$

$$S\left(t|\lambda, c\right) = \exp\left(-\lambda t^{c}\right). \tag{2}$$

Since the idea that all subjects should show the same survival patterns is an extremely restrictive assumption, we allow for differences in the rate parameter across subjects. When accounting for individual rate parameters, estimates in survival analysis may be spurious if the unobserved heterogeneity is ignored (Lancaster, 1992). On the other hand, if we address survival times using only the unobserved heterogeneity, it is impossible to specify which subjects survive longer, and why. Therefore, the individual rate parameter λ_i is assumed to be a multiplicative form of two components, λ_{1i} and λ_{2i}, which represent individual covariate effects and unobserved heterogeneity, respectively.

The first component, λ_{1i}, is represented as a linear combination of individual covariates. That is, $\lambda_{1i} = \exp\left(\mathbf{x}_i^T \boldsymbol{\gamma}\right)$, where $\mathbf{x}_i$ is a vector containing covariates related to the observable characteristics of subject i, and $\boldsymbol{\gamma}$ is a vector of all parameters associated with the covariates. Since λ_{2i} plays a role in the rate parameter as an intercept, we exclude

[4] The data from Seattle-Puget Sound were gathered from 1974; the data from Atlanta were gathered from 1975.

Table 1
Description of covariates.

Covariate	Description
age_i^j	1, if the ith subject's age ranges from $10 \times (j-1)$ to $10 \times j$, where $j = 1, \ldots, 8$ 0, otherwise
sex_i	1, if the ith subject is male 0, otherwise
$race_i^j$	1, if the ith subject is race j, where $j = 1$ (white) and 2 (black) 0, otherwise
$stage_i$	1, if the ith subject is at a stage of malignant potential 2, if the ith subject is at a stage of carcinoma in situ 3, if the ith subject is at a stage of malignant
$year_i^j$	1, if the ith subject is diagnosed in year j, where $j = 1973, \ldots, 2010$ 0, otherwise
reg_i^j	1, if the ith subject is registered in registry j, where $j = 1$ (San Francisco-Oakland), 2 (Connecticut), 3 (Detroit), 4 (Hawaii), 5 (Iowa), and 6 (New Mexico) 0, otherwise

the constant term from $\mathbf{x}_i$ in order to avoid an identification problem. The covariates comprise one continuous variable, the stage of tumor progression, and 55 dummy variables regarding age, sex, race, year of diagnosis, and registry. More details are provided in Table 1.

The second component, λ_{2i}, representing unobserved heterogeneity in the likelihood of survival, is assumed to be drawn from a gamma distribution:

$$g(\lambda_{2i}|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_{2i}^{\alpha-1} \exp(-\beta\lambda_{2i}), \tag{3}$$

where α and β are the shape and rate parameters. We use the gamma distribution both because it is the conjugate prior for the Weibull distribution, and for its flexibility. Thus, our proposed model can be summarized as follows.

$$\begin{aligned} &T_i \sim Weibull(\lambda_i, c) \\ &\lambda_i = \lambda_{1i}\lambda_{2i} \\ &\lambda_{1i} = \exp\left(\mathbf{x}_i^T \boldsymbol{\gamma}\right), \qquad \lambda_{2i} \sim Gamma(\alpha, \beta), \end{aligned} \tag{4}$$

where T_i is a random variable representing the individual survival time. By integrating the PDF over Eq. (3) given c, $\boldsymbol{\gamma}$, and λ_{2i}, the PDF and survival probability for subject i, who survives for time t_i, given c, $\boldsymbol{\gamma}$, α, and β, can be derived as

$$\begin{aligned} f(t_i|c, \boldsymbol{\gamma}, \alpha, \beta) &= \int_{\lambda_{2i}} f(t_i|c, \boldsymbol{\gamma}, \lambda_{2i})\, g(\lambda_{2i}|\alpha, \beta)\, d\lambda_{2i} \\ &= c t_i^{c-1} \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right) \left(\frac{\alpha}{\beta + t_i^c \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right)}\right) \\ &\quad \times \left(\frac{\beta}{\beta + t_i^c \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right)}\right)^\alpha \end{aligned} \tag{5}$$

$$S(t_i|c, \boldsymbol{\gamma}, \alpha, \beta) = \left(\frac{\beta}{\beta + t_i^c \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right)}\right)^\alpha. \tag{6}$$

3.2. Estimation

We estimate the parameters of the aforementioned models using maximum likelihood estimation (MLE). If a subject died before the cutoff date, his contribution to the likelihood function is the PDF of the survival time. On the other hand, for a subject who is still alive at the cutoff date, all we know is that the survival time exceeds the difference between his birth date and the cutoff date. Therefore, his contribution to the likelihood is considered to be the survival probability of the number of years before censoring occurred. Consequently, the log-likelihood function can be represented as:

$$LL = \sum_{i=1}^n \delta_i \log f(t_i) + \sum_{i=1}^n (1-\delta_i) \log S(t_i), \tag{7}$$

where t_i is the survival time and δ_i is the censoring indicator variable, which is set to one if a subject died before the cutoff date and zero otherwise.

3.3. Forecasting

In this section, we propose a method for obtaining four-year-ahead forecasts of annual numbers of cancer deaths. We begin by defining an individual probability of death, then derive it using our model. If a subject i is diagnosed in year τ_i^0, his or her probability of death, P_i^τ, can be regarded as the difference between the probabilities of surviving for more than $\tau - \tau_i^0$ years and $\tau - \tau_i^0 + 1$ years. From Eq. (6), P_i^τ can be represented as:

$$\begin{aligned} P_i^\tau &= S\left(\tau - \tau_i^0\right) - S\left(\tau - \tau_i^0 + 1\right) \\ &= \left(\frac{\beta}{\beta + \left(\tau - \tau_i^0\right)^c \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right)}\right)^\alpha \\ &\quad - \left(\frac{\beta}{\beta + \left(\tau - \tau_i^0 + 1\right)^c \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right)}\right)^\alpha. \end{aligned} \tag{8}$$

In order to obtain the above probability, we estimate the parameters c, $\boldsymbol{\gamma}$, α, and β from in-sample data using Eq. (7). The covariates of the out-of-sample subjects, $\mathbf{x}_i$, are assumed to be known, and are also used to calculate P_i^τ.

After obtaining P_i^τ from Eq. (8), it is possible to predict the annual numbers of cancer deaths. If subject i died in year τ, P_i^τ calculated from the model will be close to one. On the other hand, the probability will be close to zero in other years. Therefore, the number of cancer deaths in year τ, N_τ, can be predicted as the sum of P_i^τ, given that the subjects were diagnosed in each year τ_i^0:

$$\begin{aligned} N_\tau &= \sum_i P_i^\tau \\ &= \sum_i \left[\left(\frac{\beta}{\beta + \left(\tau - \tau_i^0\right)^c \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right)}\right)^\alpha \right. \\ &\quad \left. - \left(\frac{\beta}{\beta + \left(\tau - \tau_i^0 + 1\right)^c \exp\left(\mathbf{x}_i^T\boldsymbol{\gamma}\right)}\right)^\alpha\right]. \end{aligned} \tag{9}$$

Table 2
Estimation results.

Parameter	Estimate	(t-statistics)	Parameter	Estimate	(t-statistics)
Shape parameter (c)	1.15^{***}	(361.81)			
Covariate effect (λ_{1i})					
Age					
0–9	−0.18	(−0.17)	40–49	-1.10^{***}	(−66.17)
10–19	−0.06	(−0.24)	50–59	-1.02^{***}	(−80.50)
20–29	-3.25^{***}	(−17.04)	60–69	-0.83^{***}	(−72.08)
30–39	-1.36^{***}	(−38.87)	70–79	-0.46^{***}	(−40.68)
Sex	0.30^{***}	(47.57)			
Race					
White	0.19^{***}	(11.54)	Black	0.34^{***}	(17.95)
Year of diagnosis					
1973	0.81^{***}	(24.97)	1992	0.37^{***}	(12.81)
1974	0.71^{***}	(21.99)	1993	0.35^{***}	(12.05)
1975	0.69^{***}	(21.79)	1994	0.33^{***}	(11.48)
1976	0.62^{***}	(19.75)	1995	0.32^{***}	(11.08)
1977	0.58^{***}	(18.54)	1996	0.36^{***}	(12.60)
1978	0.51^{***}	(16.64)	1997	0.32^{***}	(11.09)
1979	0.57^{***}	(18.54)	1998	0.32^{***}	(11.03)
1980	0.51^{***}	(16.91)	1999	0.27^{***}	(9.44)
1981	0.49^{***}	(16.30)	2000	0.27^{***}	(9.51)
1982	0.47^{***}	(15.72)	2001	0.29^{***}	(10.06)
1983	0.43^{***}	(14.64)	2002	0.25^{***}	(8.74)
1984	0.46^{***}	(15.80)	2003	0.21^{***}	(7.37)
1985	0.46^{***}	(15.58)	2004	0.21^{***}	(7.23)
1986	0.42^{***}	(14.50)	2005	0.17^{***}	(6.01)
1987	0.44^{***}	(15.34)	2006	0.11^{***}	(3.77)
1988	0.43^{***}	(15.03)	2007	0.08^{***}	(2.73)
1989	0.40^{***}	(13.73)	2008	0.05	(1.60)
1990	0.37^{***}	(12.90)	2009	0.03	(0.91)
1991	0.36^{***}	(12.71)	2010	0.08^{***}	(2.78)
Registry					
San Francisco-Oakland	-0.19^{***}	(−11.02)	Hawaii	-0.17^{***}	(−7.29)
Connecticut	-0.33^{***}	(−19.05)	Iowa	-0.14^{***}	(−7.87)
Detroit	-0.18^{***}	(−10.80)	New Mexico	0.01	(0.46)
Tumor stage	0.32^{***}	(2.84)			
Unobserved heterogeneity (λ_{2i})					
Shape (α)	0.78^{***}	(108.27)			
Rate (β)	1.01^{***}	(2.65)			
Log-likelihood ($N = 396,202$)	−522,223				

*** Significant at the 1% level.

4. Results

4.1. Estimation results

Table 2 reports parameter estimates for our proposed model. Most of the coefficients are statistically significant at the 1% level, with the expected signs. Very few of the coefficients for age (<10 and 10–19) and year of diagnosis (2008 and 2009) are insignificant. This means that subjects who are aged less than 10 or between 10 and 19 have the same survival pattern as subjects over 80 years old. It can also be seen that there is no difference in the effect of the year of diagnosis among subjects diagnosed in 2008, 2009, and 2011. The fact that α and β are statistically significant is interpreted as showing that there is an unobserved heterogeneity in the rate parameter across subjects.

The shape parameter estimate is greater than one, meaning that our data show an increasing hazard rate. To validate this result, we observe the empirical hazard rates of subjects who were diagnosed in 1973, 1983, and 1993, and display the rates in Fig. 2. Fig. 2 shows that the empirical hazard rates increase over time, which is consistent with the results from our proposed model. This is also supported by the findings of Follmann and Goldberg (1988), who reported that ignoring the unobserved heterogeneity can lead to a spurious decreasing hazard rate.

We then examine the signs and magnitudes of the covariate parameter estimates in connection with the expected lifetime. The expected value of the Weibull distribution is $\lambda^{-1/c}\Gamma(1+1/c)$; this is inversely proportional to λ, given that c remains fixed. Since λ is proportional to the covariate parameter estimates $\left(\text{i.e., } \lambda = \exp\left(\mathbf{x}^T\boldsymbol{\gamma}\right)\right)$,

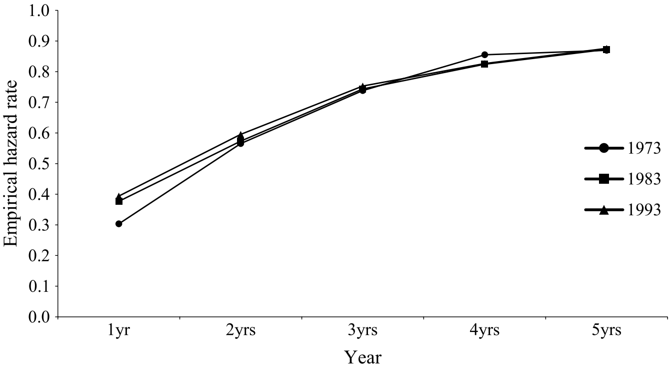


Fig. 2. Empirical hazard rates in 1973, 1983, and 1993.

the correlation between the expected lifetime and those estimates will be negative if we assume a constant unobserved heterogeneity. That is, if the covariate parameters are estimated to have positive (negative) signs, then the expected lifetime will decrease (increase). In the case of the sex covariate, for which the reference group is female, the value of 0.30 means that, on average, female subjects survive 1.30 times[5] longer than male subjects. For the race covariate, where the reference group is other races,[6] the values of 0.19 and 0.34 mean that other races survive 1.18 times longer than white subjects and 1.34 times longer than black subjects, respectively. The fact that the estimates for year of diagnosis tend to decrease as the year increases shows that subjects who were diagnosed recently will survive longer than those who were diagnosed a long time ago. This situation may arise as a result of continuous advances in medical technology.

4.2. Forecast results

To obtain four-year-ahead forecasts for 2006, 2007, 2008, 2009, 2010, and 2011, we use data until the end of 2002, 2003, 2004, 2005, 2006, and 2007, respectively. We first calculate an individual probability of death using Eq. (8), then obtain the annual number of cancer deaths using Eq. (9). As has been mentioned, we assume that the covariates of out-of-sample subjects are known, and use them to calculate the probability of death. When trying to use diagnosis year covariates in the out-of-sample period, however, there are no corresponding parameter estimates. For example, when we obtain a four-year-ahead forecast for 2006 made in 2002, estimates for the diagnosis year parameters are available only up to 2001; therefore, we must derive additional parameter estimates from 2002 to 2005. We first find an appropriate model for fitting the observable estimates (1973–2001), then extrapolate the unobservable estimates (2002–2005) using the fitted model. In order to find the proper model, we plot the fitted graphs obtained from four models: (a) $f(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \theta_3/t$, (b) $f(t) = \theta_0 + \theta_1 t + \theta_2 t^2$, (c) $f(t) = \log(\theta_0 + \theta_1 t)$, and (d) $f(t) = \exp(\theta_0 + \theta_1 t)$. Fig. 3 shows that Model (a) is better at fitting diagnosis year estimates than the other models are. We also calculate the root mean squared errors (RMSEs) of the suggested models; the RMSEs of Models (a), (b), (c), and (d) are 0.0007, 0.0018, 0.0037, and 0.0025, respectively. Therefore, we conclude that Model (a) is appropriate for extrapolating unobservable diagnosis year estimates. This result is consistent even when we change the in-sample period.

We replicate four benchmark models for the purpose of comparison. The first benchmark is the damped trend exponential smoothing method. When we compare the fifteen exponential smoothing methods from the classification described by Hyndman, Koehler, Ord, and Snyder (2008), our data are shown to follow damped trend exponential smoothing, which is equivalent to the autoregressive integrated moving average (ARIMA) (1, 1, 2) process (Hyndman et al., 2008). The second benchmark is the joinpoint regression model proposed by Kim et al. (2000). The model is fitted by least squares at a given number of change-points, called joinpoints, then the number of joinpoints is estimated. The Joinpoint software (http://surveillance.cancer.gov/join-point) shows that there are two joinpoints, at the years 1975 and 1989, and forecasts are obtained by reflecting these change points. The third benchmark is the Bayesian state space model. This model assumes that the number of cancer deaths follows a Poisson distribution, the parameter of which follows a random walk. We estimate the parameter using the Markov Chain Monte Carlo (MCMC) method. Convergence is achieved within 10,000 iterations, then an additional 90,000 iterations are used to predict the number of deaths. We pick out every 10th iteration in order to reduce the autocorrelation. Two of the models, the joinpoint regression and the Bayesian state space, were shown in a previous study (Chen et al., 2012) to perform well. The last benchmark is the prevalence, incidence, and analysis model (PIAMOD) proposed by Verdecchia et al. (2002). Unlike the other benchmarks, this one makes use of incidence data for predicting the number of deaths. The model begins by fitting an age, period, and cohort (APC) model to incidence data, then estimates the prevalence from the fitted

[5] $E_{Female}(T)/E_{Male}(T) = \left(\widehat{\lambda}_{Female}/\widehat{\lambda}_{Male}\right)^{-1/\widehat{c}} = \exp\left(-\widehat{\beta}_{Male}\right)^{-1/\widehat{c}} = \exp(-0.30)^{-1/1.15} = 1.30$.

[6] Other races includes American Indian/AK Native and Asian/Pacific Islander. For more information, see http://seer.cancer.gov/seerstat/variables/seer/race_ethnicity.

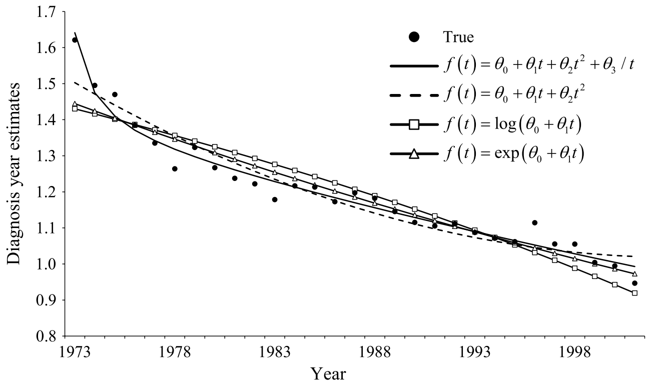


Fig. 3. Fitted lines of diagnosis year estimates, in-sample: 1973–2002.

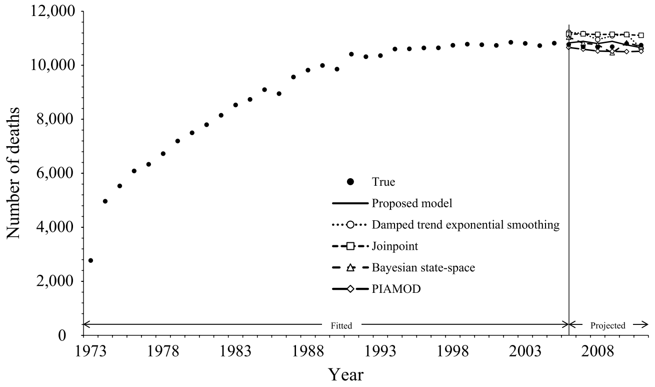


Fig. 4. Four-year-ahead forecasts of annual lung and bronchus cancer deaths for 2006–2011.

Table 3
Four-year-ahead forecast performances for 2006–2011.

Model	RMSE	MAE	MAPE (%)
Proposed model	137.8	122.6	1.15
Damped trend exponential smoothing	369.6	347.8	3.24
Joinpoint	418.8	414.9	3.87
Bayesian state space	160.3	133.7	1.25
PIAMOD	186.7	172.4	1.61

Notes: RMSE $= \frac{1}{6}\sqrt{\sum_{\tau=2006}^{2011}\left(\widehat{N}_\tau - N_\tau\right)^2}$, MAE $= \frac{1}{6}\sum_{\tau=2006}^{2011}\left|\widehat{N}_\tau - N_\tau\right|$, MAPE $= \frac{100}{6}\sum_{\tau=2006}^{2011}\frac{|\widehat{N}_\tau - N_\tau|}{N_\tau}$.

incidence and relative survival rate. The number of deaths is then derived from the fitted incidence, prevalence, and relative survival rate. APC models[7] for each period are selected based on likelihood ratio statistics, as was shown by Verdecchia, Capocaccia, Egidi, and Golini (1989).

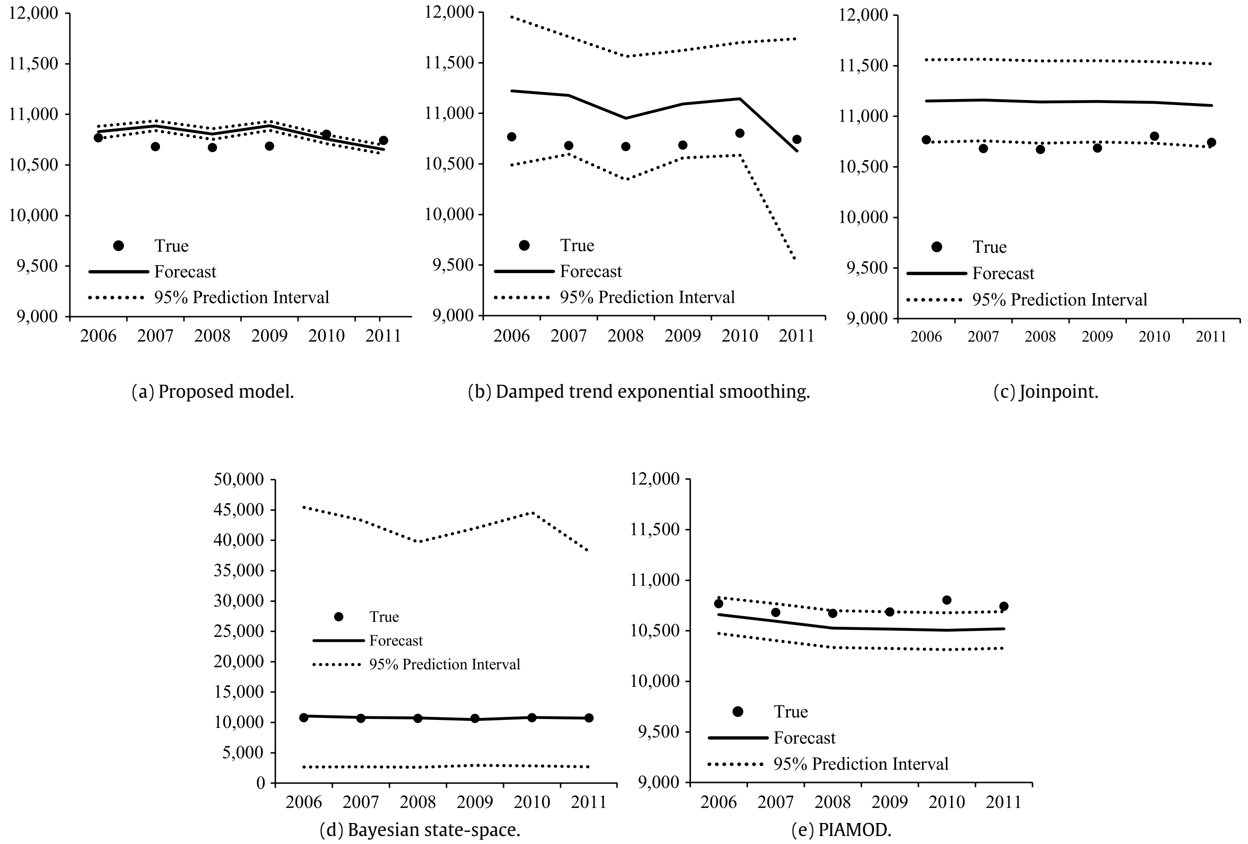
Three measurements, the root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE), are used to compare forecasting performances; the comparison results are shown in Table 3. All three measurements prove that our proposed model performs best for forecasting the annual number of cancer deaths. Among the benchmarks, the Bayesian state space model performs best. The same result can be seen in Fig. 4. Fig. 4 shows the actual and predicted numbers of cancer deaths from 2006 to 2011, comparing the numbers from our proposed model to those from the four benchmarks. While both the damped trend exponential smoothing method and the joinpoint model overestimate the actual numbers of deaths, our proposed model predicts that number very well. Although the predictions from the Bayesian state space model are closer to the actual values than are those from our proposed model for 2007 and 2008, they seem to overestimate and underestimate the numbers of deaths in 2006 and 2009, respectively.

Fig. 5 shows that our proposed model has the narrowest 95% prediction interval, being about one third of that of PIAMOD. In Fig. 5(d), the Bayesian state space has the widest prediction interval, because the parameter of the Poisson distribution is assumed to follow a random walk. This is consistent with the finding of Schmidt and Pereira (2011). All of these results emphasize that it is very important to consider both covariates and unobserved heterogeneity for survival analysis.

5. Implications

We divide the subjects into eighteen groups, characterized by race (white, black, and other race), sex (male and female), and age (50s, 60s, and 70s). We begin by estimating

[7] APC models for each period are fitted as follows (in-sample period: degree of age/period/cohort): 1973–2002: 5/0/2, 1973–2003: 7/0/4, 1973–2004: 7/0/4, 1973–2005: 7/0/4, 1973–2006: 7/0/6, 1973–2007: 7/0/6.



(a) Proposed model.

(b) Damped trend exponential smoothing.

(c) Joinpoint.

(d) Bayesian state-space.

(e) PIAMOD.

Fig. 5. 95% prediction intervals of annual lung and bronchus cancer deaths for 2006–2011.

the effect of the diagnosis year on the survival time for each group. The estimation results for the individual groups are not consistent with those from the whole data set; as can be seen in Table 4, not all groups show significant diagnosis year effects. Interestingly, there are hardly any diagnosis year effects for groups 7 to 18. This means that, for blacks and other subjects, there are no differences in survival times between subjects who were diagnosed recently and those diagnosed a long time ago. This result suggests that advances in medical technology have only influenced the survival times of white subjects who belong to groups 1 to 6. In addition, there are also differences within these six groups. In the case of group 1, white male subjects in their fifties, there are significant diagnosis year effects just until 1976. On the other hand, groups 2 and 3, white male subjects in their sixties and seventies, show significant effects until 2001 and 1998, respectively. The aging of populations could be one explanation for this phenomenon.

Next, we derive the probability of surviving more than $t\,(t = 1, 2, 3)$ years for the subjects in each group who are diagnosed in the latest year that we can observe. Suppose that T_g is a random variable representing the survival times of group $g\,(g = 1, 2, \ldots, 18)$. From Eq. (6), their survival probabilities are computed as:

$$S\left(t|\widehat{c}_g, \widehat{\boldsymbol{\gamma}}_g, \widehat{\alpha}_g, \widehat{\beta}_g\right) = \left(\frac{\widehat{\beta}_g}{\widehat{\beta}_g + t^{\widehat{c}_g}\overline{\lambda_{1g}}}\right)^{\widehat{\alpha}_g}, \tag{10}$$

where $\widehat{\alpha}_g$ and $\widehat{\beta}_g$ are the estimated shape and rate parameters of the gamma distribution in each group, and $\widehat{c}_g$ is the estimated shape parameter of the Weibull distribution. $\overline{\lambda_{1g}}$ is defined as $\left(1/N_g\right)\sum_{i=1}^{N_g} \exp\left(\mathbf{x}_i^T \widehat{\boldsymbol{\gamma}}_g\right)$, where N_g is the size of group g and $\widehat{\boldsymbol{\gamma}}_g$ is an estimated vector of the regression parameters. Fig. 6 shows the probabilities of surviving for more than 1, 2, and 3 years for each group. Not only do we see that white subjects are likely to survive longer than black subjects, we also find that female subjects are likely to survive longer than male subjects. Furthermore, it is observed that, overall, older subjects have shorter survival times. These results are consistent with the estimation results in Table 2.

6. Concluding remarks

In this paper, we propose a Weibull mixture model with individual covariates and unobserved heterogeneity in order to examine the effects of demographic variables on individual survival times, and to predict the

Table 4
Heterogeneity in diagnosis year effects for each group.

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Age	50	60	70	50	60	70	50	60	70	50	60	70	50	60	70	50	60	70
Sex	Male	Male	Male	Female	Female	Female	Male	Male	Male	Female	Female	Female	Male	Male	Male	Female	Female	Female
Race	White	White	White	White	White	White	Black	Black	Black	Black	Black	Black	Other	Other	Other	Other	Other	Other
Size	33,662	67,001	68,188	22,776	41,795	44,638	6,464	9,058	6,651	3,500	4,686	3,864	2,519	4,809	5,206	1,522	2,540	2,867
LL	−49,011	−90,022	−75,360	−37,863	−65,311	−58,929	−8,491	−11,282	−6,953	−5,276	−6,646	−4,621	−3,811	−6,728	−6,216	−2,488	−4,003	−3,815
1973	0.52	0.78	0.72	0.77	0.93	0.63	0.67	0.58	0.36	0.85	0.85	1.20	0.64	0.63	0.58	1.02	0.95	0.09
1974	0.43	0.70	0.59	0.63	0.76	0.79	0.14	0.40	0.25	0.80	0.80	0.82	0.81	0.02	0.54	1.22	0.24	1.26
1975	0.38	0.77	0.54	0.59	0.75	0.55	0.53	0.21	0.38	0.21	0.64	0.60	0.92	0.69	0.50	0.52	0.94	1.21
1976	0.24	0.56	0.47	0.55	0.78	0.67	0.16	0.53	−0.08	0.73	0.73	0.70	1.09	0.52	−0.03	1.25	0.53	1.28
1977	0.11	0.54	0.53	0.53	0.72	0.68	0.42	0.23	0.18	0.71	0.40	0.37	1.40	0.48	0.14	0.96	0.35	1.51
1978	0.15	0.46	0.43	0.40	0.58	0.51	0.18	0.24	0.19	0.58	0.31	0.49	0.52	0.77	0.40	0.68	−0.04	0.94
1979	0.16	0.53	0.52	0.48	0.55	0.64	0.37	0.27	0.24	0.36	0.22	0.59	0.60	0.65	−0.11	1.23	1.93	1.03
1980	0.08	0.40	0.47	0.42	0.57	0.65	0.09	0.29	0.15	0.35	0.54	0.23	0.34	0.56	0.48	1.34	0.63	1.13
1981	0.08	0.42	0.49	0.44	0.61	0.62	0.23	0.26	0.05	0.05	0.56	0.33	0.24	0.48	−0.11	0.75	0.87	0.72
1982	0.02	0.38	0.43	0.41	0.60	0.57	0.17	−0.01	0.42	0.53	0.56	0.35	0.37	0.33	0.25	1.30	0.60	0.41
1983	0.03	0.41	0.38	0.36	0.43	0.65	0.17	0.04	0.04	0.46	0.40	0.22	0.60	0.12	0.48	0.81	0.72	0.25
1984	0.07	0.44	0.35	0.31	0.60	0.65	0.07	0.10	−0.01	0.43	0.62	0.68	0.46	0.42	0.71	1.11	0.85	0.10
1985	−0.04	0.42	0.37	0.31	0.53	0.59	0.21	0.13	0.41	0.53	0.65	0.19	0.96	0.09	0.12	1.64	0.40	0.66
1986	0.07	0.35	0.31	0.48	0.48	0.58	0.23	0.14	0.04	0.29	0.46	0.45	0.25	0.18	0.15	0.89	0.82	1.06
1987	0.12	0.31	0.36	0.39	0.54	0.64	−0.03	0.10	0.08	0.27	0.63	0.47	0.30	0.20	0.38	1.29	0.51	0.68
1988	0.08	0.35	0.29	0.39	0.57	0.54	0.01	0.23	0.20	0.59	0.39	0.41	0.87	0.10	0.12	1.88	0.16	0.84
1989	0.13	0.34	0.26	0.45	0.45	0.47	0.32	0.03	0.28	0.26	0.63	0.42	0.50	0.03	−0.17	1.34	0.56	1.32
1990	0.06	0.25	0.29	0.32	0.42	0.49	0.17	−0.03	0.20	0.30	0.45	0.75	1.08	0.34	0.28	1.03	0.47	0.69
1991	−0.08	0.26	0.23	0.38	0.45	0.51	0.45	0.09	0.05	0.33	0.73	0.22	0.25	0.35	−0.24	1.74	0.33	1.08
1992	−0.02	0.28	0.21	0.30	0.47	0.47	0.35	0.14	0.31	0.42	0.73	0.50	0.64	0.26	0.15	1.03	0.00	0.42
1993	−0.11	0.35	0.24	0.33	0.40	0.43	0.16	0.15	−0.06	0.50	0.56	0.13	0.67	0.34	0.08	0.90	0.47	0.90
1994	0.03	0.27	0.18	0.27	0.40	0.38	0.10	0.20	0.09	0.64	0.04	0.47	0.40	0.50	0.07	1.41	0.97	0.95
1995	0.02	0.28	0.23	0.29	0.36	0.44	0.11	0.02	−0.13	0.59	0.37	0.24	0.35	0.12	0.07	0.49	0.33	0.85
1996	0.05	0.25	0.21	0.32	0.55	0.39	0.04	−0.10	0.30	0.56	0.63	0.43	0.54	0.05	0.17	1.04	−0.19	0.74
1997	0.01	0.19	0.17	0.22	0.42	0.41	0.23	0.04	−0.02	0.51	0.41	0.26	0.18	0.19	−0.08	0.97	0.70	0.64
1998	0.01	0.21	0.20	0.21	0.35	0.36	0.44	0.14	0.12	−0.09	0.66	0.34	0.83	0.22	0.05	0.76	0.31	1.05
1999	−0.03	0.19	0.11	0.02	0.38	0.44	0.00	0.14	−0.12	0.19	0.30	0.13	0.74	−0.18	0.08	0.36	0.35	0.54
2000	−0.09	0.27	0.12	0.12	0.36	0.43	0.02	0.05	−0.07	0.19	0.24	0.58	−0.05	0.13	0.02	0.65	0.37	0.55

(continued on next page)

Table 4 (*continued*)

Group	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Age	50	60	70	50	60	70	50	60	70	50	60	70	50	60	70	50	60	70
Sex	Male	Male	Male	Female	Female	Female	Male	Male	Male	Female	Female	Female	Male	Male	Male	Female	Female	Female
Race	White	White	White	White	White	White	Black	Black	Black	Black	Black	Black	Other	Other	Other	Other	Other	Other
Size	33,662	67,001	68,188	22,776	41,795	44,638	6,464	9,058	6,651	3,500	4,686	3,864	2,519	4,809	5,206	1,522	2,540	2,867
LL	−49,011	−90,022	−75,360	−37,863	−65,311	−58,929	−8,491	−11,282	−6,953	−5,276	−6,646	−4,621	−3,811	−6,728	−6,216	−2,488	−4,003	−3,815
2001	−0.12	0.28	0.27	0.13	0.22	0.35	0.10	0.34	0.02	0.56	0.12	0.44	0.52	−0.39	0.30	0.44	0.38	0.78
2002	0.11	0.15	0.16	0.05	0.26	0.29	0.11	0.18	0.21	0.17	0.28	0.31	0.27	0.30	0.20	0.77	−0.13	0.67
2003	0.00	0.11	0.11	0.14	0.19	0.32	−0.06	0.24	0.20	0.58	0.03	0.33	0.33	0.15	0.17	0.81	−0.07	0.37
2004	−0.04	0.19	0.10	0.18	0.18	0.24	0.18	0.09	−0.23	−0.17	0.72	0.49	0.39	−0.20	0.16	1.08	0.15	0.51
2005	−0.14	0.21	0.06	0.23	0.11	0.21	0.06	0.08	0.04	0.09	0.07	0.01	0.65	0.16	−0.03	0.54	−0.08	0.60
2006	−0.11	0.02	−0.03	0.04	0.14	0.09	−0.10	−0.28	0.02	0.26	0.11	0.58	1.03	0.06	−0.19	0.62	0.43	0.48
2007	−0.14	0.12	−0.02	−0.08	0.03	0.21	0.18	0.02	−0.11	−0.26	0.12	−0.09	0.16	0.30	−0.22	1.29	0.01	0.07
2008	−0.18	−0.03	−0.13	−0.03	0.08	0.10	−0.07	0.07	−0.32	0.26	0.04	0.23	0.34	−0.08	0.26	0.41	−0.15	0.75
2009	−0.10	−0.03	−0.13	0.10	−0.01	0.12	0.17	0.06	−0.20	−0.11	0.09	0.11	0.56	−0.10	−0.39	0.68	−0.43	0.32
2010	−0.16	0.05	−0.06	0.12	−0.03	0.13	−0.23	0.16	0.04	−0.01	0.22	0.02	0.57	−0.13	0.05	0.18	−0.08	0.41

Notes. The estimates in the shaded cells are significant at the 0.05 level. LL denotes log-likelihood.

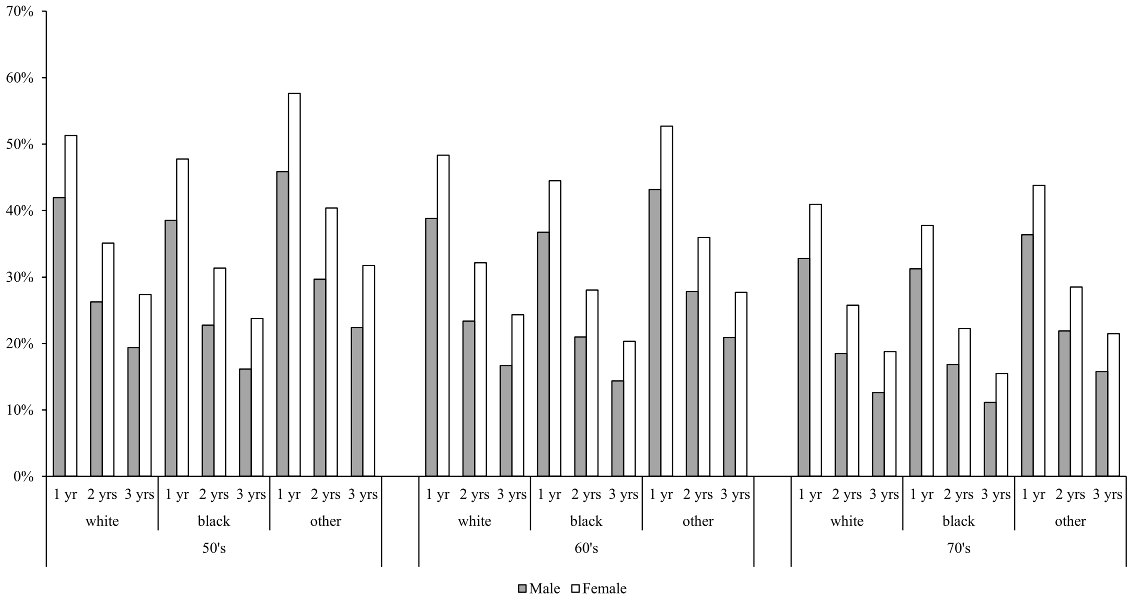


Fig. 6. Probabilities of surviving for more than 1, 2, and 3 years for each group. *Note.* The survival probabilities are calculated from 2011.

annual number of cancer deaths by adopting a bottom-up strategy. A comparison of our results with those from the four benchmarks – the damped trend exponential smoothing method, the joinpoint regression model, a Bayesian state space model, and PIAMOD – reveals that our proposed model performs best for predicting annual numbers of cancer deaths. Furthermore, by segmenting the patients based on age, sex, and race, we are able to specify the differences between groups, and assess the group-specific survival probabilities within a given period. Our results show that older, female, and white patients survive significantly longer than younger, male, and black patients, respectively. Also, patients diagnosed in recent years survive significantly longer than those diagnosed a long time ago.

Despite the demonstrated predictive performance of our proposed model, this study has two limitations. First, when predicting the annual number of deaths, we assumed that the incidence of out-of-sample subjects, including covariates, is known. If we did not allow our model to use these data, the prediction accuracy would decrease. Second, unlike the benchmarks, our model requires a great deal of data, including individual information. Though individual data are often highly accessible in developed countries, such might not be the case in some developing countries. In such countries, the applicability of our model is restricted. However, as Parkin (2006) mentioned, cancer registration has come to be the norm over the last 60 years, and it seems reasonable to expect future expansion in both the geographic coverage and the scope of work. The exponential growth observed in computing power will also make our model more efficient in dealing with large amounts of data.

Nevertheless, our study is distinctive in its adoption of a bottom-up strategy for predicting aggregate-level units. This differs from the method of predicting the number of deaths using only a time series model, as our proposed model shows how aggregate numbers can be obtained from individual-level units. Since we made use of a larger amount of individual information, our model enables us not only to specify the differences between groups based on age, sex, and race, but also to obtain more precise results than are possible when using the other benchmarks. We believe that our study can shed light on important issues in both the public and private sectors. For governments, it may be beneficial in enabling health and welfare budgets to be set in a more precise way. By utilizing predictions such as the four-year-ahead annual numbers of cancer deaths, governments will be able to hedge financial risks. With regard to the private sector, our segmentation results can provide guidance to insurance firms, allowing them both to target existing customers more efficiently and to attract new customers by supplying customized cancer insurance products.

Acknowledgments

The authors thank Hyune-Ju Kim for introducing this interesting research topic to us, and also thank anonymous reviewers for their valuable comments and suggestions.

References

American Cancer Society (2011). *Global cancer facts and figures* (2nd ed.). Atlanta: American Cancer Society.

Chen, H. S., Portier, K., Ghosh, K., Naishadham, D., Kim, H. J., Zhu, L., et al. (2012). Predicting US- and state-level cancer counts for the current calendar year. *Cancer*, *118*, 1091–1099.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *34*(2), 187–220.

Duchateau, L., & Janssen, P. (2007). *The frailty model*. Springer.

Follmann, D. A., & Goldberg, M. S. (1988). Distinguishing heterogeneity from decreasing hazard rates. *Technometrics*, *30*(4), 389–396.

Hakulinen, T., & Tenkanen, L. (1987). Regression analysis of relative survival rates. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *36*(3), 309–317.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London, Series A (Mathematical, Physical and Engineering Sciences)*, *454*, 903–995.

Hyndman, R. J., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer.

Kim, H. J., Fay, M. P., Feuer, E. J., & Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, *19*(3), 335–351.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, *47*(4), 939–956.

Lancaster, T. (1992). *The econometric analysis of transition data*. Cambridge: Cambridge University Press.

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

Parkin, D. M. (2006). The evolution of the population-based cancer registry. *Nature Reviews Cancer*, *6*, 603–612.

Prentice, R. L., & Kalbfleisch, J. D. (1979). Hazard rate models with covariates. *Biometrics*, *35*(1), 25–39.

Royston, P., & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, *21*, 2175–2197.

Schmidt, A. M., & Pereira, J. B. M. (2011). Modelling time series of counts in epidemiology. *International Statistical Review*, *79*(1), 48–69.

Siegel, R., Jiemin, M., Zou, Z., & Jemal, A. (2014). Cancer statistics, 2014. *CA: A Cancer Journal for Clinicians*, *64*(1), 9–29.

Tiwari, R. C., Ghosh, K., Jemal, A., Hachey, M., Ward, E., Thun, M. J., & Feuer, E. J. (2004). A new method of predicting US and state-level cancer mortality counts for the current calendar year. *CA: A Cancer Journal for Clinicians*, *54*(1), 30–40.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*, 439–454.

Verdecchia, A., Capocaccia, R., Egidi, V., & Golini, A. (1989). A method for the estimation of chronic disease morbidity and trends from mortality data. *Statistics in Medicine*, *8*, 201–216.

Verdecchia, A., De Angelis, R., & Capocaccia, R. (2002). Estimation and projections of cancer prevalence from cancer registry data. *Statistics in Medicine*, *21*, 3511–3526.

Duk Bin Jun is a Professor at KAIST College of Business. He received his doctorate from the U. C. Berkeley. He has published in the *International Journal of Forecasting, Journal of Forecasting, Journal of Business and Economic Statistics, Technological Forecasting and Social Change, Marketing Letters, Telecommunications Policy, Telecommunication Systems*, and other journals. His research interests include adaptive forecasting, structural changes in time series analysis, business cycle forecasting, new product diffusion and choice processes, and telecommunication forecasting.

Kyunghoon Kim is a doctoral student at KAIST College of Business. His research interests include survival analysis in the healthcare industry.

Myoung Hwan Park is a Professor in the Department of Industrial Engineering, Hansung University. He received his doctorate from KAIST in 1993. His research interests include telecommunications forecasting, new product forecasting and supply chain management. He has published in the *International Journal of Forecasting, Telecommunication Systems, Computers and Operations Research*, and other journals.