Communications for Statistical Applications and Methods
2021, Vol. 28, No. 2, 119–134

https://doi.org/10.29220/CSAM.2021.28.2.119
Print ISSN 2287-7843 / Online ISSN 2383-4757

Study of age specific lung cancer mortality trends in the US using functional data analysis

Bhikhari Tharu[1,a], Keshav Pokhrel[b], Gokarna Aryal[c], Ram C. Kafle[d], Netra Khanal[e]

[a]Department of Mathematics, Spelman College, USA;
[b]Department of Mathematics and Statistics, University of Michigan-Dearborn, USA;
[c]Department of Mathematics and Statistics, Purdue University Northwest, USA;
[d]Department of Mathematics and Statistics, Sam Houston State University, USA;
[e]Department of Mathematics, University of Tampa, USA

Abstract

Lung cancer is one of the leading causes of cancer deaths in the world. Investigation of mortality rates is pivotal to adequately understand the determinants causing this disease, allocate public health resources, and apply different control measures. Our study aims to analyze and forecast age-specific US lung cancer mortality trends. We report functions of mortality rates for different age groups by incorporating functional principal component analysis to understand the underlying mortality trend with respect to time. The mortality rates of lung cancer have been higher in men than in women. These rates have been decreasing for all age groups since 1990 in men. The same pattern is observed for women since 2000 except for the age group 85 and above. No significant changes in mortality rates in lower age groups have been reported for both gender. Lung cancer mortality rates for males are relatively higher than females. Ten-year predictions of mortality rates depict a continuous decline for both gender with no apparent change for lower age groups (below 40).

Keywords: functional data analysis, lung cancer, mortality rate, principal component analysis, basis function, time series

1. Introduction

Lung cancer is the most common cancer in males and the third most common in females mortality rates in the world (Ferlay *et al.*, 2019). In the United States, lung cancer is the cause of death for 24% of all cancer deaths, and 13% of all new cancer diagnoses (Thun, 2008; Howlader *et al.*, 2019). It is the leading cause of US cancer deaths, regardless of gender and ethnicity (Thun, 2008; Howlader *et al.*, 2019; Siegel *et al.*, 2019). Almost as many Americans die of lung cancer every year than that of the prostate, breast, and colon cancer combined. The 5-year survival rate in the United States for lung cancer is 15.5%. Despite the advancements in medical treatment during the last decades, advances in lung cancer survival are yet to be achieved compared to other malignant cancers. Developing countries get around 50% of the Global cases, whereas, in 1980, 69% of the cases were in developed countries. In 1964, the US public health service published a seminal report causally relating smoking to lung cancer. World Health Organization (WHO) predicts there will be around 1.9 billion smokers by 2025 with a sufficient increase in tobacco consumption, especially in Asia (Dela *et al.*, 2011).

[1] Corresponding author: Department of Mathematics, Spelman College, 350 Spelman Lane, SW, Science Center, 330 - Box: 953, Atlanta, GA 30314, USA. E-mail: btharu@spelman.edu

Published 31 March 2021 / journal homepage: http://csam.or.kr

It is believed that the first comprehensive study of lung cancer was carried out by Adler in 1912. Many studies show that smoking is the major cause of lung cancer. At the beginning of 1900, the average smoker used to smoke less than 100 cigarettes per year, whereas in 1960, it increased to 3600 cigarettes per year. Multiple case-control studies have been carried out relating different kinds of smoking and lung cancer (Boffetta *et al.*, 1999). In this report, however, we focus on studying mortality rates and their prediction. The accurate estimation and projection of mortality rates is essential for planning and managing the public health policies of an identified population to allocate resources. There are significant differences in time trends of mortality rates of males and females, as well as age-specific trends that may be distinctive to separate countries and their populations. Establishing parameters of the geographic population by country and further narrowing the predictions by gender of each population is essential to better understand the disease dynamics. This study is restricted to addressing the geographic population of the US and further distinguishing males from females within that population.

Numerous studies have been conducted to analyze US lung cancer mortality. It has been widely analyzed by using an age-period, age-cohort, age-period-cohort model, including frequentist and the Bayesian setting (Takahashi *et al.*, 2001; Rosenberg and Anderson, 2011; Tharu *et al.*, 2015; Smith and Wakefield, 2016). The time trends of lung cancer mortality rates have been projected using different approaches such as the generalized additive model, differential equation model, the age-period model, and the joinpoint regression method (Negri *et al.*, 1990; Clements *et al.*, 2005; McCarthy *et al.*, 2012; Kafle *et al.*, 2014; John and Hanke, 2016). However, the past attempts to model and forecast the difference between males and females lung cancer mortality rates have not been sufficiently addressed. These attempts have only considered the most recent information to forecast the mortality rates and hence may not be able to capture the long term trends of the data. A different approach to modeling such data with minimal assumptions and forecasting is therefore required, one that forecasts by incorporating a large scale time series data.

In this article, we have implemented a functional time series model where the functional principal component (Ramsay, 2005) decomposes the smooth curves to the basis functions. The method is used to model and forecast the lung cancer mortality trends for different age-groups and gender. This method differs from previous methods in the following ways. (i) It does not assume that the data follows a particular distribution but assumes it is from an underlying smooth function. (ii) It uses the entire age mortality curves to forecast and does not rely solely on recent observations. Forecasting performance suggests that the functional data analysis approach increases the accuracy of predictions (Booth *et al.*, 2006). The main focuses of this article are to - study the time trend of age-specific mortality rates for US males and females. - predict age-specific lung cancer mortality rates for US males and females. - compare the time trend of age-specific lung cancer mortality rates within the group.

2. Material and method

2.1. Material

Annual age-specific mortality rates of lung cancer data for US males and females from 1969 to 2017 are obtained from the National Cancer Institute (NCI) (SEER, 2019). The data is extracted from the NCI database using SEER*Stat software. The NCI provides mortality rates for 5-year age groups, but some age groups have an insufficient number of cases. We have regrouped the available data into 10-year intervals: 20–30, 30–40, 40–50, 50–60, 60–70, 70–80, and above 80 years. These age groups are represented by 25, 35, 45, 55, 65, 75, 85 respectively thereafter. The age group 20–30 for females is not included in the study because of a large number of missing values. The annual mortality rates are reported using the census population of the first day of July as a proxy for person-years at risk for each calendar year.

2.2. Methodology

Annual lung cancer mortality of males and females are considered as a function of ages. The mortality rate is defined as the number of deaths per 100,000 in a year. Recognizing that there is not a universal transformation technique that can be applied for all types of data to reduce out of sample variance and forecast error (Hyndman and Booth, 2008), this study took the log transformation of the data. The analysis is performed to the transformed data ($y_t(x)$). The functional time series model assumes the underlying smooth function of the data with some observed error. We have implemented a nonparametric smoothing technique (Ferraty and Vieu, 2006) to the transformed data $y_t(x) = \log(y_t^*(x))$ to prepare a smooth curve ($s_t(x)$) as functional data object. This study adapts the penalized regression splines smoothing (Wood, 1994) method. Since the choice of a smoothing parameter is crucial, the generalized cross-validation (GCV) criteria has been adapted to find the optimal tuning parameter λ (Craven and Wahba, 1979).

GCV is defined as:

$$\mathrm{GCV}(\lambda) = \left(\frac{n}{n - \mathrm{df}(\lambda)}\right)\left(\frac{\mathrm{SSE}}{n - \mathrm{df}(\lambda)}\right),$$

where SSE represents the sum of squares due to error. The right factor of GCV is the unbiased estimate of error variance σ^2 similar to regression analysis, and thus represents some discounting by $\mathrm{df}(\lambda)$ from n, $\mathrm{df}(\cdot)$ represents the degrees of freedom. The left factor further discounts this estimate by multiplying by $(n/(n - \mathrm{df}(\lambda)))$ (Ramsay and Silverman, 2005). In some practical cases GCV depicts a tendency of under smoothing the data. Therefore, Gu and Kim (2008) suggest to multiply it by factors such as 1.2 or 1.4 and ensure that the additional discounting does not seriously increase the odds of over-smoothing the data. We now have the functional time series of $(x_i, y_t(x_i))$, $t = 1, 2, 3, \ldots, n$, $i = 1, 2, 3, \ldots, m$ and

$$y_t(x_i) = s_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i}, \tag{2.1}$$

where $y_t(x_i)$ is the lung cancer mortality rate for age group x_i in year t, $\varepsilon_{t,i}$ is an independent and identically distributed standard normal variates with the amount of error $\sigma_t(x_i)$ that varies on age and time. We obtained "m" smooth functional curves as functional data objects (Erbas *et al.*, 2007; Hyndman and Ullah, 2007) for the analysis by:

$$s_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x), \tag{2.2}$$

where $e_t(x)$ is the uncorrelated error of the model, $\mu(x)$ is the mean curve of $s_t(x)$ across years, $\{\phi_k(x)\}$ is a set of orthonormal basis functions estimated using principal component decomposition method (Ramsay and Dalzell, 1991), and K represents the optimal number of principal components in the model.

Functional principal component (FPC) is applied to the smooth curves $\{s_t(x_i)\}$ which provides the minimum number of basis functions, enables informative interpretation, and gives coefficients $\beta_{t,k}$ which are uncorrelated with each other (Erbas *et al.*, 2007). The coefficients $\beta_{t,k}$ are obtained by using a univariate time series model, autoregressive integrated moving average (ARIMA). Four basis functions have provided a reasonable fit using a functional regression model where the first two basis functions explain most of the variability in the data. This method's efficacy is that it identifies the small number of basis functions and therefore simplifies interpretations and creates uncorrelated coefficients. The basis function ($\phi_k(x)$) is modeled as a weight assigned to the variable which takes the maximum and the minimum values at the highest and the lowest peak of the curve. A model is chosen based on the minimum mean integrated squared error (MISE) for the basis function with a given number of K. For a given K, and selected basis functions $\phi_k(x)$ the mean integrated squared error (MISE) can be defined as:

$$\text{MISE} = \frac{1}{n}\sum_{t=1}^{n}\int e_t^2(x)dx.$$

We used robust method to obtain $\mu(x)$ proposed by Hyndman and Ullah, 2007. The data in our study do not contain any visible outliers or any other unusual behaviour, so we estimate $\mu(x)$ using the mean of $s_t(x)$ over t, and apply functional principal component decomposition. The FPCs are computed by first constructing the $q \times n$ matrix G with $(j,t)^{th}$ element $s_t(x^*) - \hat{\mu}(x_j^*)$ where $\{x_1^*, x_2^*, \dots, x_q^*\}$ is fine grid of equally spaced values that span the interval $[x_1, x_q]$. In addition, the singular value decomposition of G gives $G = \Phi\Lambda V$, where $\hat{\phi}_k(x_j^*)$ is the $(j,k)^{th}$ element of Φ and $\hat{\beta}_{t,k}$ is the $(t,k)^{th}$ element of $G'\Psi$. Please refer to (Hyndman and Ullah, 2007) for further details about the estimation procedures.

2.3. Forecast the mortality rates

The univariate time series model are fitted to each coefficients $\beta_{t,k}$, $t = 1, \dots, n$ and these estimates are used to find the coefficients $\beta_{t,k}$, $t = n+1, \dots, n+h$ and $k = 1, \dots, K$. The basis functions ϕ_k are obtained by using functional principal component method and the coefficients are obtained by using univariate time series method. The coefficients $\beta_{t,k}$ and $\beta_{t,l}$ are assumed to be uncorrelated for $k \neq l$. Therefore, univariate method will be an adequate method to forecast each time series $\hat{\beta}_{t,k}$. We use forecast coefficients with Equation (2.2) to obtain $s_t(x)$, $t = T+1, \dots, T+h$. From Equation (2.2), forecast of $s_t(x)$ are also the forecasts of $y_t(x)$ (Hyndman and Ullah, 2007).

Combining Equations (2.1) and (2.2), we can write

$$y_t(x_i) = \mu(x_i) + \sum_{k=1}^{K}\beta_{t,k}\phi_k(x) + e_t(x) + \sigma_t(x_i)\varepsilon_{t,i}. \tag{2.3}$$

This implies

$$\eta_{T,h} = \text{Var}\left[y_{T+h}|y_{t(x_i)}\right] \approx \hat{\sigma}_\mu^2(x) + \sum_{k=1}^{K}\text{Var}\left(\beta_{T+h}|\beta_{1,k}, \dots, \beta_{T,k}\right)\hat{\phi}_k^2(x) + \text{var}(e_t(x)) + \sigma_{T+h}^2(x).$$

In addition, the variance of $(\beta_{T+h}|\beta_{1,k}, \dots, \beta_{T,h})$ can be obtained from the time series model, the variance of smooth estimate $\hat{\mu}(x)$ can be obtained from the smoothing method, observational error variance is obtained by assuming binomial distribution of mortality rates, and the model error variance is obtained by averaging $\hat{e}_t{}^2(x)$ for each x. Assuming the errors are normally distributed, a $100(1-\alpha)\%$ prediction interval is constructed as $\hat{y}_{T,h}(x) \pm z_{\alpha/2}\sqrt{\eta_{T,h}}$.

The exponential smoothing state-space model (Hyndman *et al.*, 2008) is then used to forecast log-transferred mortality rates and prediction intervals (Erbas *et al.*, 2007). The mean integrated squared

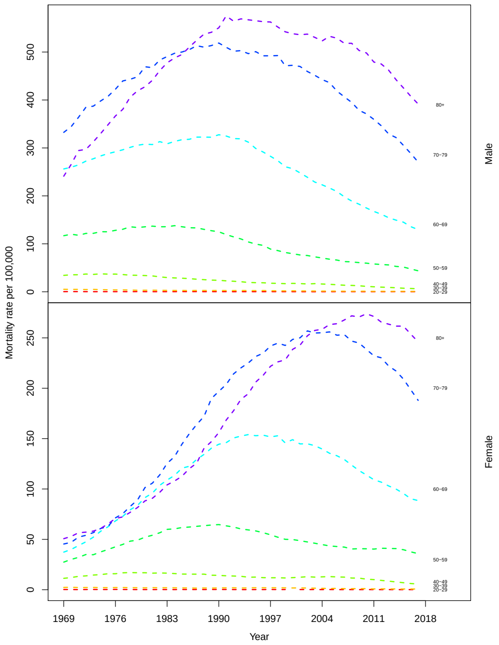


Figure 1: *Lung cancer mortality rates per 100,000 for US males (top) and females (bottom) from 1969–2017.*

forecasting error is then used to evaluate the accuracy of the estimated predictions for the estimates. We estimated a 10-year prediction (2018–2027) of the lung cancer mortality for males and females for each age group. Part of the results obtained for this paper has been facilitated by the statistical package “ftsa” (Hyndman, 2019). Sample Rcodes are available in an Appendix A1.

3. Results and discussion

Figure 1 displays the observed lung cancer mortality rates within the US for both males (top chart) and females (bottom chart) from 1969 to 2017. The mortality rates of all the male age groups have continued to decline since 1990 (Siegel *et al.*, 2019), and the same trend has been observed for the females below age 70. But the decline in mortality for older females has been more pronounced and began somewhat later: after 2000 for age group 75, and after 2010 for the age group 85. The male lung cancer mortality rates are higher than that of females in all age groups. Higher mortality rates for males indicate that they are more at risk of lung cancer than females (Patel *et al.*, 2004; Tabatabai *et al.*, 2016). The mortality rates for males and females below the age of 40 remained at essentially the same level. This may raise questions about new treatment procedures developed during the last few decades. Rates for males and females above the age of 70 declined at a higher rate since 2010 (John and Hanke, 2016; DeSantis *et al.*, 2019). Although the global trend shows that the number of cancer deaths is increasing, individual death rates for the US are falling (Roser and Ritchie, 2020).

In Figure 2, the top panel represents the observed and the smooth log-mortality rates per 100,000 males while the bottom panel represents the females. We have incorporated the penalized regression splines (Wood, 1994) for smoothing the log mortality rates of the observed data. The colors in the figures represent a chronicle order of the rainbow starting at the oldest 1969 as red, and the most recent

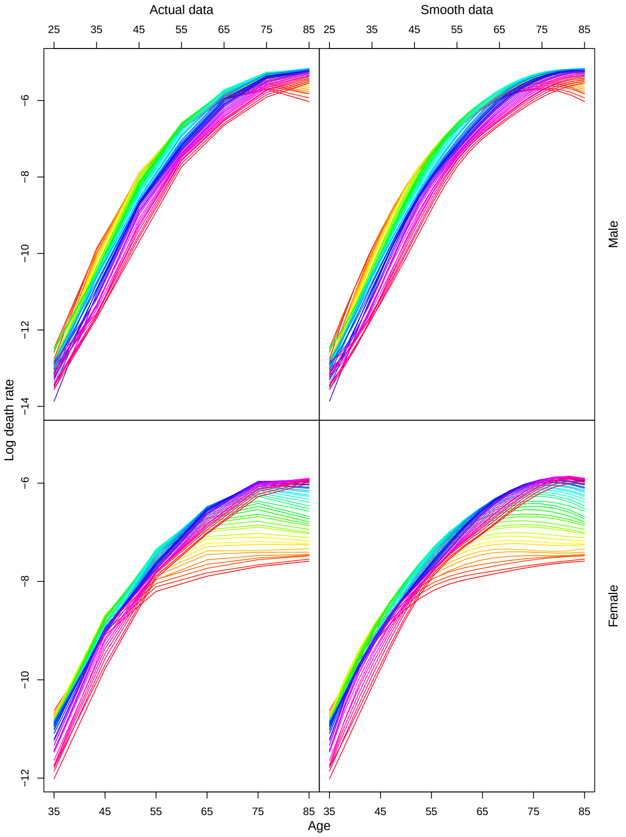
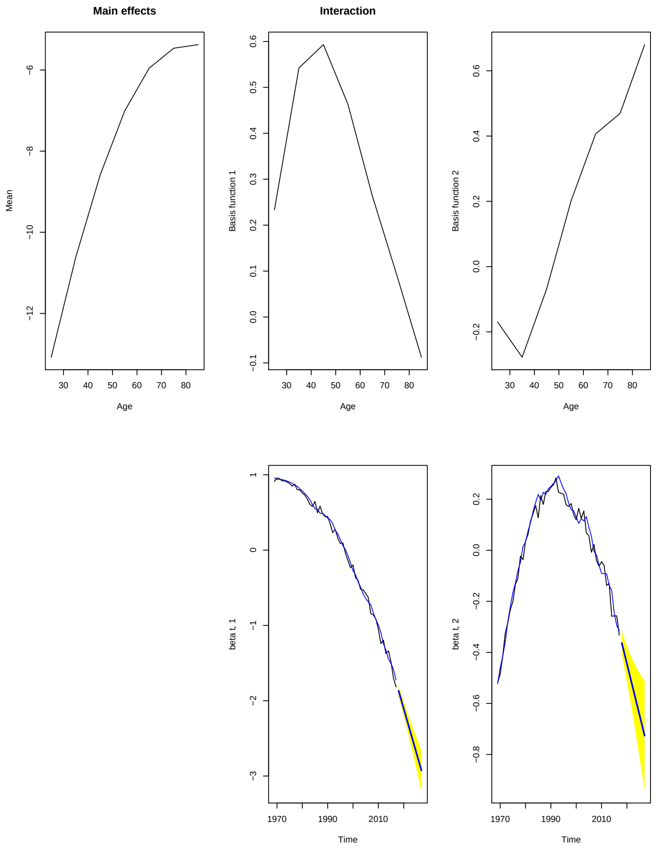


Figure 2: *The left panel represents observed log-mortality rates for males and females while the smooth log-mortality rates are in the right panel from 1969–2017. Rainbow color code: red to purple from 1969–2017.*

2017 as purple. These figures depict that female mortality rates have higher variability than males; however, the mortality rates are higher in males than females. The mortality rates for both gender are decreasing in recent years.

Figure 3 shows the first two functional principal components and their associated scores (in black color) with 80% prediction intervals (in yellow color) using an exponential smoothing state-space model for the US male population. From left to right, the first graph represents the estimated mean curve, the second and third graphs represent basis functions, the fourth and fifth represent coefficient corresponding to Basis function 1 and Basis function 2 respectively. In the bottom panel, the black curves represent estimated coefficients by the functional principal component method and the blue curves are estimated through the ARIMA model. The first two basis functions explain 82.52% and 13.95% variability of the data. The Basis function 1, $\phi_1(x)$ models males at ages around 45, and the Basis function 2, $\phi_2(x)$ models males at ages around 35. The decrease in $\beta_{t,1}(x)$ and $\beta_{t,2}(x)$ accelerates during the last 20 years.

Figure 4 is for female population. For females, $\beta_{t,1}(x)$ models at age around 70 and $\beta_{t,2}(x)$ models between 60 to 70. The coefficients $\beta_{t,1}(x)$ and $\beta_{t,2}(x)$ show a decreasing trend in the past 20 years, however $\beta_{t,2}$ has different trend during 1980–1995. Two basis functions for females explain 78.08%


Main effects

Interaction

Figure 3: *The first two principal components and associated scores of lung cancer mortality rates for the males.*

and 14.97% variability of the data. Since forecast accuracy is significantly affected by outliers, bivariate and functional highest density region (HDR) boxplot are used to identify the outliers (Hyndman and Shang, 2010). We did not notice potential outliers in the data.

We selected optimal number of principal components $K = 2$ because (i) the first two principal components explain sufficiently large variation (more than 90%) of the subject response variable, (ii) adding an extra principal component merely adds 0.5% of the variability of the data, (iii) the interpretation of the third and the fourth principal components is unclear other than referring it to unexplained variations of the first two principal components.

The rainbow plot in Figure 5 (top) displays the forecast of male lung cancer mortality rate for 2018–2027 using the exponential smoothing state-space model (Hyndman *et al.*, 2008) whereas the bottom panel displays for the females; the gray color represents the data used for estimation. This rainbow plot contains curves that are ordered chronologically within the rainbow – the year (2018) is red and the year (2027) is purple. In order to enhance the forecast accuracy of the model, sufficiently large number of principal components $K = 4$ has been chosen in the study (Hyndman and Booth, 2008). However, two principal components provide every indication of adequacy for the analysis. Mortality rates increase with increasing age; nonetheless the mortality rates are greater during recent

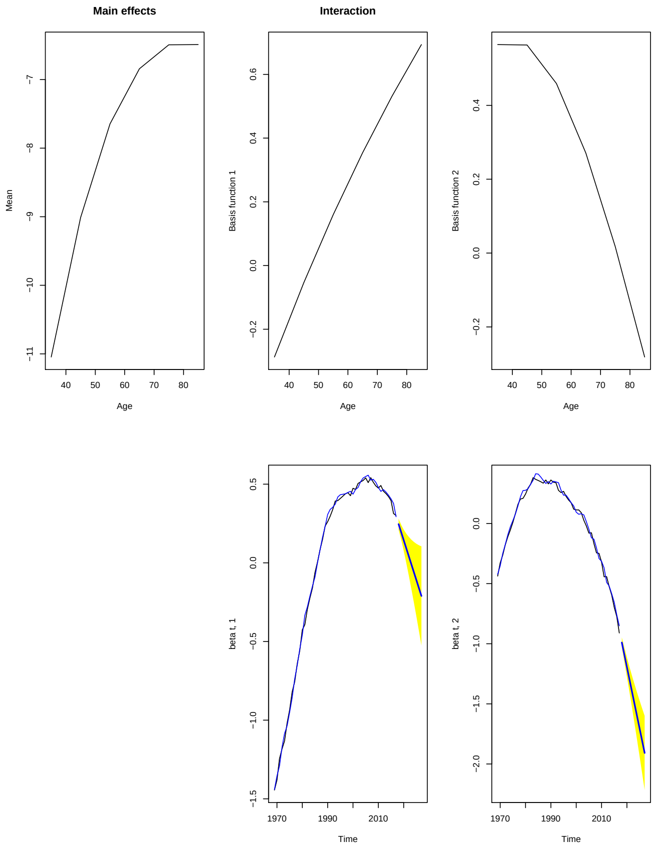


Figure 4: *The first two principal components and associated scores of lung cancer mortality rates for the females.*

years than what was seen in the past for the older age groups (above 70, in particular for females). Cancer patients of ages above 80 have the highest mortality rates accompanied by a lower survival rate. This group is likely to have cancer at higher stages, less likely to have complex surgery, functional declines, and undertreatment (DeSantis *et al.*, 2019). The predicted mortality rates appear to decrease for all age groups regardless of gender; however, the rate is slightly higher in females between ages 60 to 70. Our results are consistent with the results found by a number of studies such as Alberg *et al.* (2013), Siegel *et al.* (2020), Kazerouni *et al.* (2004), and Tabatabai *et al.* (2016). This decrease could be due to tobacco control efforts implemented since the 1960s, the advances in detection and diagnosis, and improvements in the treatments (Kazerouni *et al.*, 2004; Jeon *et al.*, 2018; Rapp *et al.*, 1988; Corrales, 2018). In addition, immunotherapy and chemotherapy play a significant role in reducing lung cancer mortality in the USA and the world.

Table 1 presents the forecast of expected mortality rates per 100,000 and 80% prediction interval by gender and age groups: 20–30, 30–40, 40–50, 50–60, 70–80, and above 80 for the year 2018, 2022, and 2027. The mortality rates are expected to decrease across the age groups by an average of 34% for males and 38% for females between 2018 and 2027. However, the differences in predicted mortality rates for specific age groups are more pronounced. In particular, the decrease in mortality rates of young adults (below 45) females is at least 10% higher than that of males.

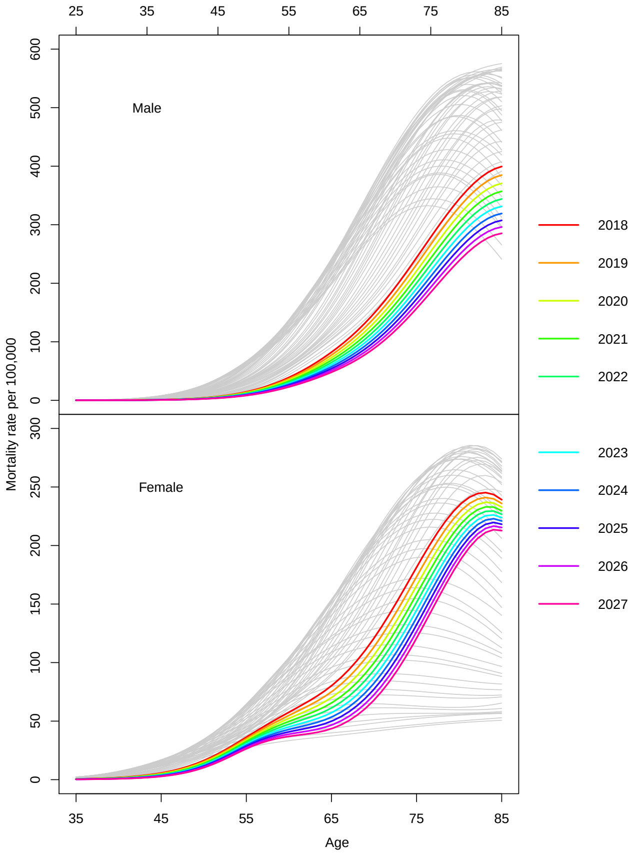


Figure 5: *Age-specific lung cancer mortality rates for males (top) and females (bottom) as functional time series for all age groups from 1969 to 2017 (gray color). The predicted mortality rates from 2018 to 2027 are presented using red to purple color in the order of the rainbow.*

Table 1: Predicted lung cancer mortality rates per 100,000 of US males and females of each age-specific group for the years 2018, 2022, and 2027 with 80% prediction interval

	Age	2018			2022			2027		
		Lower	Mean	Upper	Lower	Mean	Upper	Lower	Mean	Upper
Male	20–30	0.115	0.159	0.218	0.100	0.150	0.224	0.086	0.140	0.229
	30–40	0.724	0.791	0.864	0.580	0.648	0.724	0.423	0.505	0.602
	40–50	6.100	6.636	7.218	4.654	5.146	5.689	3.235	3.744	4.333
	50–60	37.020	39.132	41.364	28.872	30.834	32.929	20.793	22.891	25.199
	60–70	119.662	123.169	126.777	93.514	97.259	101.152	67.492	72.395	77.652
	70–80	269.303	278.843	288.723	220.068	232.129	244.851	166.200	184.583	204.999
	above 80	383.785	399.503	415.864	323.097	344.010	366.275	252.099	285.354	322.995
Female	30–40	0.541	0.606	0.678	0.368	0.443	0.533	0.202	0.288	0.413
	40–50	5.443	5.723	9.016	3.788	4.120	4.481	2.320	2.775	3.306
	50–60	35.126	36.274	37.460	29.540	31.509	33.610	22.725	26.393	30.654
	60–70	76.726	80.168	83.764	57.673	61.205	64.952	38.889	43.527	48.718
	70–80	172.226	181.418	191.100	142.224	151.422	161.216	108.612	120.850	134.470
	above 80	226.089	239.158	252.982	210.878	226.735	243.785	186.164	212.734	243.095

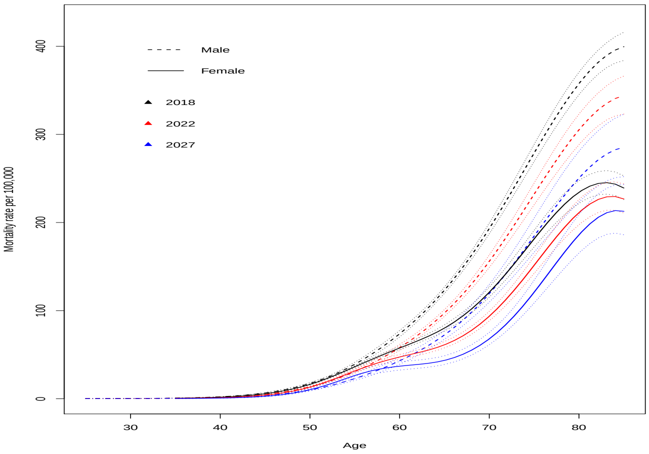


Figure 6: *Predicted lung cancer mortality rates of males and females. Males are represented in dashed lines and females are in solid lines. Black, red, and blue colors represent the years 2018, 2022, and 2027 respectively with corresponding 80% prediction intervals in dotted lines.*

Figure 6 presents predictions of mortality rates for males and females for three different years (black-2018, red-2022, and blue-2027) along with 80% prediction intervals. A continuous decline in mortality rates for both gender is apparent. Furthermore, we expect fewer female fatalities than males in the future (Table 1). Since 80% prediction intervals for males and females are narrower (within a margin of error of 30 to 50 counts per 100,000), the model appears to capture expected mortality rates and predicts the data well. The elderly population (more than 60 years) show the consistently higher prediction of mortality rates for males compared to females (Table 1 and Figure 6).

Figure 7 depicts age-specific mortality rates per 100,000, a ten-year prediction, and 80% prediction intervals of the estimates for males and females. The forecast shows that lung cancer mortality rates are expected to decrease across all the age groups with no visible difference in trends.

We observed higher declining mortality rates for ages 60 to 80 years, which suggests a possible influence of recent increases in health awareness (Siegel, 2020; DeSantis *et al.*, 2019). The mortality rates for the age group below 30 for males and females are likely to be stable (John and Hanke, 2016) in the future because of advanced treatments and early detection (Devesa *et al.*, 1989; Rapp *et al.*, 1988; Corrales *et al.*, 2018; Antonia *et al.*, 2014) as well as strong body functionality. The forecast suggests that mortality is higher in males than in females in each age group. Some of the major reasons for the higher mortality rates for males are likely to be due to weaker responses to the treatments, lower likelihood to survive at different stages of lung cancer, lower immune power against the effects of smoking, environmental adaptability, and the difference in lifestyles (Pauk *et al.*, 2005). It is also observed that women respond to some of the chemotherapy treatments better than men. Similarly, the surgical treatments of lung cancer benefit women more than men at all stages of the disease (Tabatabai *et al.*, 2016). In addition, females are more susceptible than males to carcinogenic substances (Patel *et al.*, 2004). The wider prediction intervals exhibited at the end of each age-specific groups in the ten-year forecast are indicators of greater uncertainty for the long term prediction (Figures 7). The model does come with some limitations that include: (i) the effect of the birth cohort, which has not been included in the study, despite being a significant factor that impacts mortality trends, (ii) the smoothing method adapted in this report may reduce the natural variability present in the data.

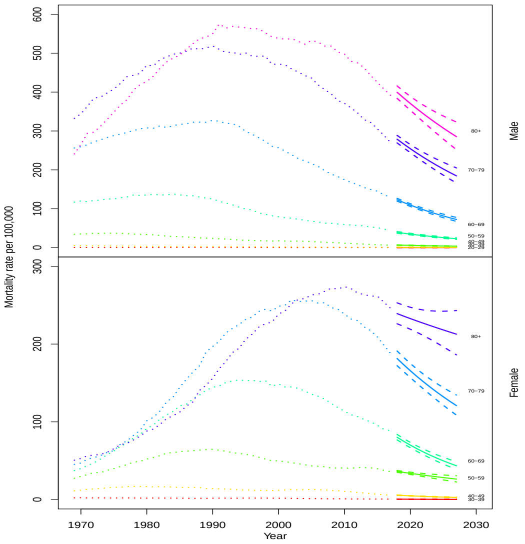


Figure 7: *Observed data (1969-2017) with an estimated ten-year prediction (2018-2027) and 80% prediction intervals of age-specific mortality rates of males (top) and females (bottom).*

The inconsistency in predicted mean lines with the observed data in the oldest age groups: 70-80, and 80+ is noticeable. In the oldest age group women outnumber men because of their longer life expectancy; in 2016, there were 4.2 million women aged 85 years and older compared with 2.2 million men, or 186 women for every 100 men. The number of male patients in these age groups is reasonably less than that of women. This also creates higher variability in mortality rates and results with wider prediction intervals of mortality rates. The lifetime carcinogenic exposures, somatic mutations, and age-related changes in immune system also adds up to mortality rates (DeSantis *et al.*, 2019). Most importantly, reasons for subsequent decline in risk with higher variability in mortality rates and inconsistencies in predicted mean lines in those age groups are unclear.

Cancer in 80+ age group appears in more advanced stages. The patients with cancer in this age group have the lowest relative survival of any age group and largest disparities noted when cancer is diagnosed at advanced stages. In addition, these patients are less likely to receive surgical treatment. Treating patients aged 85 years and older who have cancer is complex because of the higher likelihood of comorbid conditions, declines in health status associated with aging, and dearth of data regarding treatment procedure in this age group. The census bureau reported that the United States population has continued gray (DeSantis *et al.*, 2019). Elderly population is one of the fastest growing age group in the United states. There is an increasing need of a comprehensive study of evidence-based treatment procedure to develop treatment guide for these age groups.

The proposed model is under the functional paradigm, direct comparisons with non-functional models are not appropriate. To compare the functional and non-functional models, the last five years of data (2013–2017) are used for validation. We have fitted the models using mortality rates from 1969 to 2012 and predicted for the last five years. The different models are compared using the sum of squares due to errors (SSE). The lower the sum of the square due to error, the better the prediction. The proposed model outperforms the generalized Poisson regression and the generalized additive models with various spline procedures. The sum of squares due to errors in Poisson regression for males is 43117.1 and for females 35955. However, the generalized additive model with cubic regression splines is competitive with the proposed model. The SSE using proposed models for males and females are 10093.5 and 1737.7 whereas the generalized additive model with cubic spline produced SSE of 10908.28 and 1203.22 for males and females respectively. The mean integrated sum of squares due to errors for the males are 0.02713 and 0.6757 for the females.

4. Conclusion

Understanding the trend of mortality rates with accurate prediction is imperative for disease prevention and public health planning. In addition, the effects of demographic and spatial attributable variables will create much-needed attention for the prevention and control of the disease to the particular clusters of the population. We applied a novel method, the functional time series analysis, in order to model US lung cancer mortality rates of males and females. Even though we observed a declining trend in mortality, lung cancer (38.5) shows a stark difference in mortality rates (per 100,000) compared to other highly vulnerable cancers such as prostate(19.0), colon and rectum (13.7), breast (11.1), pancreatic(11.0) cancer (SEER, 2019). The developed model displays the strengths of good fit and better prediction accuracy. We consistently observed lower prediction error by functional time series model compared to other competitive statistical methods. In addition, the basis functions capture the main features and trends by taking historical data into consideration. The penalized spline smoothing method applied to the log-transformation of the observed lung cancer mortality rates is flexible to model sudden change in trend, missing data, and long term prediction. The proposed modeling procedure can be adapted to model and predict public health, environmental, and financial indices to name a few. Accurate prediction of mortality trends is fundamental to create awareness, design public health policies, develop proper etiological epidemiology, and diagnosis procedures. The following are the highlights of some of the notable features of this study:

- The trends of lung cancer mortality rates of the US males and females are different. Mortality rates are higher for males than in females in every age group.
- Ten-year prediction for the US lung cancer mortality rates suggests a continuing decline for both males and females. However, the mortality rates for young adults (below 40) are stable during the period of study.
- The combination of functional time series and the decomposition of the principal component analysis method is better than existing competitive methodologies to model and forecast the mortality rates.

The current study has not incorporated the effects of smoking, birth cohort, and seasonal impacts. By incorporating these variables could lead to an interesting advancement of the current study.

Acknowledgements

The authors would like to thank three anonymous reviewers for their valuable feedback and comments which significantly improved the paper's quality.

Declaration of Interest

We confirm that there are no known conflicts of interest associated with this publication and there has been no financial support for this work that could have influenced its outcomes.

Appendix:

Sample codes

```
# library(ftsa)
```

Nonparametric smoothing to prepare functional data

```
# sm_maledata<-as.list(rep(NA,7))
# id<-seq(25,85,10)
# for(i in 1:length(id)){
#   sm_male<-male[male[,1]==id[i],]
#   sm_out_male<-min.np(sm_male$trate1,
#                       type.S =S.KNN,
#                  par.CV = list(criteria="GCV"))
#   sm_out1_male<-sm_out_male$fdata.est
#   sm_maledata[[i]][1:49]<-sm_out1_male$data
# }
```

Basis functions

```
# M.set<-c(4,5,6)
# r.set<-c(2,3,4)
# ini.method="EM"
# basis.method="bs"
# sl.v=rep(0.5,10)
# max.step=50
# grid.l=seq(0,1,0.01)
# grids=seq(0,1,0.002)
# result<-fpca.mle(sm_malematrix, M.set,r.set,ini.method,
#                  basis.method,sl.v,max.step,grid.l,grids)
# summary(result)
# M<-result$selected_model[1]
# r<-result$selected_model[2]
```

Forecasting smoothed data

```
# male<-as.data.frame(sm_matrix)
# spread_data<-spread(male,time,value)
# spread_data<-as.matrix(spread_data)
# age<-unique(male[,1])
# fds_data<-spread_data[,-1]
# ages<-c("6","7","8","9","10","11","12",
#         "13","14","15","16","17","18")
# rownames(fds_data)<-ages
#
```

```
# time<-unique(male[,3])
# time<-as.ts(time)
# x<-age
# y<-fds_data
# new_sm<-list(x=x,y=y,time=time)
# plot(forecast(ftsm(new_sm,order=2),h=10),"components")
```

Checking the ourtliers in the data

```
# functional bag plot for outliers #
# outl_male<-sfts(ts(as.numeric(smooth_male1$y),frequency = 7))
# fboxplot(data = outl_male, plot.type = "functional",
#          type = "bag", projmethod="PCAproj")
# fboxplot(data = outl_male, plot.type = "bivariate",
#          type = "bag", projmethod="PCAproj")
```

References

Alberg AJ, Brock MV, Ford JG, Samet JM, and Spivack SD (2013). Epidemiology of Lung Cancer: Epidemiology of Lung Cancer Diagnosis and Management of Lung Cancer, 3rd ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines, *Chest*, **123**, 21-49.

Antonia SJ, Brahmer JR, Gettinger S, *et al.* (2014). Nivolumab (Anti–PD–1; BMS–936558, ONO–4538) in Combination With Platinum-Based Doublet Chemotherapy (PT-DC) in Advanced Non-Small Cell Lung Cancer (NSCLC): Metastatic Non-Small Cell Lung Cancer, *International Journal of Radiation Oncology*Biology*Physics*, **90**, S2.

Boffetta P, Nyberg F, Agudo A, *et al.* (1999). Risk of lung cancer from exposure to environmental tobacco smoke from cigars, cigarillos and pipes, *International Journal of Cancer*, **6**, 805–806.

Booth H, Hyndman RJ, Tickle L, and Jong PD (2006). Lee-Carter mortality forecasting, *Demographic Research*, **15**, 289–310.

Clements MS, Armstrong BK, and Moolgavkar SH (2005). Lung cancer rate predictions using generalized additive models, *Biostatistics*, **6**, 576–589.

Corrales L, Scilla K, Caglevic C, Miller K, Oliveira J, and Rolfo C (2018). Immunotherapy in Lung Cancer: A New Age in Cancer Treatment, *Advances Experimental Medical Biology*, **995**, 65–95.

Craven P and Wahba G (1979). Smoothing noisy data with spline functions, *Numerische Mathematik*, **31**, 377–403.

Dela Cruz CS, Tanoue LT, and Matthay RA (2011). Lung Cancer: Epidemiology, Etiology, and Prevention, *Clinical Chest Medicine*, **32**, 605–644.

DeSantis CE, Miller KD, Dale W, *et al.* (2019). Cancer statistics for adults aged 85 years and older, *CA: A Cancer Journal for Clinicians*, **69**, 452-467.

Devesa SS, Blot WJ, and Fraumeni Jr. JF (1989). Declining Lung Cancer Rates Among Young Men and Women in the United States: A Cohort Analysis, *JNCI: Journal of the National Cancer Institute,* **81**, 1568–1571.

Erbas B, Hyndman RJ, and Gertig DM (2007). Forecasting age-specific breast cancer mortality using functional data models, *Statistics in Medicine*, **26**, 458-470.

Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, Znaor A, and Bray F (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods, *International Journal of Cancer*, **144**, 1941–1953.

Ferraty F and Vieu P (2006). *Nonparametric Functional Data Analysis*, Springer-Verlag, New York.

Gu C and Kim Y (2008). Penalized likelihood regression: General formulation and efficient approximation, *The Canadian Journal of Statistics*, **30**, 619–628.

Howlader CK, Noone AM, and Krapcho M *et al.* (2019). SEER Cancer Statistics Review: 1975–2016, *National Cancer Institute*, Bethesda, from : https://seer.cancer.gov/csr/1975–2016/, based on November 2018 SEER data submission, posted to the SEER web site, April.

Hyndman RJ (2019). https://cran.r-project.org/web/packages/ftsa/citation.html

Hyndman RJ and Booth H (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration, *International Journal of Forecasting*, **24**, 32–342.

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008). *Forecasting with Exponential Smoothing: a State Space Approach*, Berlin: Springer.

Hyndman RJ and Shang HL (2010). Rainbow Plots, Bagplots, and Boxplots for Functional Data, *Journal of Computational and Graphical Statistics*, **19**, 29–45.

Hyndman RJ and Ullah MS (2007). Robust forecasting of mortality and fertility rates: A functional data approach, *Computational Statistics Data Analysis*, **51**, 4942–4956.

Jeon J, Holford TR, Levy DT, *et al.* (2018). Smoking and Lung Cancer Mortality in the United States From 2015 to 2065: A Comparative Modeling Approach, *Annals of Internal Medicine*, **169**, 684–693.

John U and Hanke M (2016). Age and Sex-Specific Trends in Lung Cancer Mortality over 62 Years in a Nation with a Low Effort in Cancer Prevention, *International Journal of Environmental Research and Public Health*, **13**, 362.

Kafle RC, Khanal N, and Tsokos CP (2014). Bayesian age-stratified joinpoint regression model: an application to lung and brain cancer mortality, *Journal of Applied Statistics*, **41**, 2727–2742.

Kazerouni N, Alverson CJ, Redd SC, Redd SC, Mott JA, and Mannino DM (2004). Sex Differences in COPD and Lung Cancer Mortality Trends—United States, 1968–1999, *Journal of Women's Health*, **13**, 17–23.

McCarthy WJ, Meza R, Jeon J, and Moolgavkar SH (2012). Lung Cancer in Never Smokers: Epidemiology and Risk Prediction Models, *Risk Analysis*, Chapter 6 , **32**, 69–84.

Negri E, La Vecchia C, Levi F, Randriamiharisoa A, Decarli A, and Boyle P (1990). The application of age, period and cohort models to predict Swiss cancer mortality, *Journal of Cancer Research and Clinical Oncology*, **116**, 207–214.

Patel JD, Bach PB, and Kris MG (2004). Lung Cancer in US Women: A Contemporary Epidemic, *Journal of the American Medical Association*, **291**, 1763–1768.

Pauk N, Kubík A, *et al.* (2005). Lung cancer in women, *Lung Cancer*, **48**, 1–9.

Ramsay JO and Dalzell CJ (1991). Some Tools for Functional Data Analysis, *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**, 539–561.

Ramsay JO and Silverman BW (2005). Functional Data Analysis, *Springer*, New York.

Rapp E, Pater JL, Willan A, *et al.* (1988). Chemotherapy can prolong survival in patients with advanced non-small-cell lung cancer report of a Canadian multicenter randomized trial, *Journal of Clinical Oncology*, **6**, 633–641.

Rosenberg PS and Anderson WF (2011). Age-Period-Cohort Models in Cancer Surveillance Research: Ready for Prime Time?, *Cancer Epidemiology Biomarkers Prevention*, **20**, 1263–1268.

Roser M and Ritchie H (2020). Cancer, *Our world in data*.

SEER (2019). Surveillance, Epidemiology, and End Results (SEER) Program Research Data (1975–2016), National Cancer Institute, DCCPS, *Surveillance Research Program based on the November 2018 submission*.

Siegel RL, Miller KD, and Jemal A (2019). Cancer statistics, *CA: A Cancer Journal for Clinicians*, **69**, 7-34.

Siegel RL, Miller KD, and Jemal A (2020). Cancer statistics, *CA: A Cancer Journal for Clinicians*, **70**, 7–30.

Smith TR and Wakefield J (2016). A Review and Comparison of Age-Period-Cohort Models for Cancer Incidence, *Statistical Science*, **31**, 591–610.

Tabatabai MA, Kengwoung-Keumo JJ, Oates GR, Guemmegne JT, Akinlawon A, Ekadi G, Fouad MN, and Singh KP (2016). Racial and Gender Disparities in Incidence of Lung and Bronchus Cancer in the United States: A Longitudinal Analysis, *PloS One*, **11**.

Takahashi H, Okada M, and Kano K (2001). Age-Period-Cohort Analysis of Lung Cancer Mortality in Japan, 1960–1995, *Journal of Epidemiology*, **11**, 151–159.

Tharu BP, Kafle RC, and Tsokos CP (2015). Bayesian age-period-cohort model of lung cancer mortality, *Epidemiology Biostatistics and Public Health*, **12**.

Thun MJ (2008). Lung Cancer Occurrence in Never-Smokers: An Analysis of 13 Cohorts and 22 Cancer Registry Studies, *PLOS Medicine*, **5**.

Wood SN (1994). Monotonic Smoothing Splines Fitted by Cross Validation, *SIAM Journal on Scientific Computing*, **15**, 1126–1133.

Received June 19, 2020; Revised October 18, 2020; Accepted January 28, 2021