Journal of Data Science **8**(2010), 339-348

Age-Adjusted US Cancer Death Rate Predictions

Matthew J. Hayat[1], Ram C. Tiwari[2], Kaushik Ghosh[3], Mark Hachey[4], Ben Hankey[5] and Rocky Feuer[5]

[1]*Johns Hopkins University,* [2]*National Cancer Institute,* [3]*University of Nevada Las Vegas,* [4]*Information Management Services and* [5]*National Cancer Institute*

Abstract: The likelihood of developing cancer during one's lifetime is approximately one in two for men and one in three for women in the United States. Cancer is the second-leading cause of death and accounts for one in every four deaths. Evidence-based policy planning and decision making by cancer researchers and public health administrators are best accomplished with up-to-date age-adjusted site-specific cancer death rates. Because of the 3-year lag in reporting, forecasting methodology is employed here to estimate the current year's rates based on complete observed death data up through three years prior to the current year. The authors expand the State Space Model (SSM) statistical methodology currently in use by the American Cancer Society (ACS) to predict age-adjusted cancer death rates for the current year. These predictions are compared with those from the previous Proc Forecast ACS method and results suggest the expanded SSM performs well.

Key words: Age-adjusted mortality rate, local quadratic model, state space model.

1. Introduction

This year, more than 1,500 people a day in the United States are expected to die of cancer (American Cancer Society, 2007). Accounting for one in every four deaths, cancer is the second-leading cause of death, exceeded only by heart disease. Cancer is a major public health problem and estimates of up-to- date age-adjusted cancer death rates are desired by researchers and public health administrators involved in the war on cancer because of the need to make accurate assessments of progress being made. However, collecting mortality data nationwide results in a three-year time lag in reporting mortality statistics which stems from the time required to collect and process mortality data from all states and report mortality for individual cancers. It is also important to note that additional

resources could possibly shorten the time lag somewhat, but could not reduce it to zero. An alternative approach involving forecasting methodology is considered here and offers the possibility of obtaining reasonably accurate up-to-date cancer mortality rates.

Each year, the American Cancer Society publishes the estimated number of new cancer cases and deaths for the current calendar year based on projections from observed data available through the most recent year in its publication *Cancer Facts & Figures*. Studying age-adjusted cancer death rates is of interest since cancer death counts do not account for the size of the population. The purpose of this paper is to extend this paradigm to age-adjusted cancer death rates. The methodology is intended to project an updated rate based on historical data through three years prior to the current year. Statistical methodology using the State Space Model is proposed that adjusts for short term trends at the end of the utilized data range. However, this methodology does not consider factors that may influence rate changes or recent trends. Age-adjusted cancer death rates are projected to the current year (2007) based on data collected since 1969.

2. Cancer Mortality Data

Data on deaths in the United States are compiled by the National Center for Health Statistics (NCHS) of the Centers for Disease Control and Prevention (CDC)[1]. Cause of death is based on reasons cited on the death certificate. The process of data collection, compilation, and publication takes several years, causing up to a three year time lag in the death data available to the public. For example, in the current year 2007, the most recent national level death file available includes data from 1969 to 2004. The Surveillance, Epidemiology and End Results (SEER) program annually obtains from NCHS a public-use file containing information on all deaths occurring in the US by calendar year[2]. Information on each death includes age at death, sex, geographic area of residence, and underlying and contributing causes of death. The underlying cause of death is used in the calculation of age-adjusted death rates. Cause of death before 1999 was coded according to ICD-9; beginning with deaths in 1999, ICD-10 was used (World Health Organization, 1975, 1992). Age-adjusted death rates for the SEER geographic areas, for each state, and for the entire US are obtained using SEER*Stat software available[3].

[1]See http://www.cdc.gov/nchs/Default.htm
[2]See http://seer.cancer.gov/mortality
[3]See http://seer.cancer.gov

3. Materials and Methods

3.1 Death rate prediction methods

The American Cancer Society (ACS) has used two different methods for age-adjusted death count projections to the current year (Wingo *et al.*, 1998; Pickle *et al.*, 2003). The first method was an extension of the forecasting methodology for age-adjusted cancer death count projections between 1995 and 2003. The second method, which is currently in use, is based on the State Space Model (SSM) (Tiwari *et al.*, 2004; Ghosh *et al.*, 2007). For our analysis, we have extended this method to project the age-adjusted death rates, in place of age-adjusted death counts, for the current year to overcome the three year time lag in data availability.

Between 1995 and 2003, the American Cancer Society used the PROC FORECAST procedure in the SAS software system (henceforth denoted by PF) to project the age-adjusted cancer death counts (Ghosh and Tiwari, 2007). This method gives three-year-ahead predictions and 95% prediction intervals for the age-adjusted death counts. This model was easily adaptable to rates by simply replacing the age- adjusted death counts with rates without changing the form of the model. We explain the PF method in the following paragraph.

The age-adjusted mortality rate at time t is defined as

$$r_t = \sum_{j=1}^{J} w_j \frac{d_{tj}}{n_{tj}},$$

where w_j are the known standards (weights) normalized to sum to 1 and d_{tj} and n_{tj} are the number of deaths and population at risk at time t and in age-group j. Note that there are $J = 19$ age-groups in the SEER Program given by $0-1,\ 1-4,\ 5-9,\ \ldots,\ 85+$. The PF method assumes that the age-adjusted death rates have a quadratic trend with autoregressive errors given by

$$r_t = b_0 + b_1 t + b_2 t^2 + u_t,$$

where $u_t = a_1 u_{t-1} + \cdots + a_p u_{t-p} + \epsilon_t$ are the autoregressive errors (Wingo *et al.*, 1998). Here $\{\epsilon_t\}$ is assumed to be an independent sequence of zero-mean, random errors with constant variance. Model fitting occurs in two sequential steps. First, the least-squares method is used to estimate the trend parameters $\hat{b}_0,\ \hat{b}_1,\ \hat{b}_2$. Then, an autoregressive model is fit on the residuals from this estimated model $\hat{u}_t = r_t - \hat{b}_0 - \hat{b}_1 t - \hat{b}_2 t^2$. The final model so obtained was used to obtain 3-year ahead predictions and the corresponding 95% prediction intervals. The results of the PF presented here are the point estimate predictions obtained from PF.

Tiwari *et al.* (2004) developed an alternative method for projecting age-adjusted cancer death counts to the current year. Motivated to improve sensitivity to recent short term trends and eliminate subjectivity, these authors used a state space model method (SSM) for predicting the age-adjusted death counts. This model is currently used by the American Cancer Society for estimating current year counts.

The SSM is easily adapted for predicting age-adjusted death rates, with the model for r_t written as (measurement equation)

$$r_t = \alpha_t + \epsilon_t, \qquad (t = 1,\ 2,\ \ldots),$$

where α_t is the unobserved trend and ϵ_t is the (measurement) error at time t. The ϵ_t's are assumed to be serially uncorrelated with mean 0 and constant variance σ_t^2, independent of time t. Instead of using a deterministic function in PF to model the trend, we follow the framework of Tiwari *et al.* (2004) and use a local quadratic trend that changes with time. This allows the model to quickly make adjustments and get closer to the observed series (Ghosh *et al.*, 2007). The form of the local quadratic time-varying trend (for $t = 1,\ 2,\ \ldots$) is the transition equation as follows:

$$\begin{cases} \alpha_t &= \alpha_{t-1} + \beta_{t-1} + \gamma_{t-1} + \eta_{1t}, \\ \beta_t &= \beta_{t-1} + 2\gamma_{t-1} + \eta_{2t}, \\ \gamma_t &= \gamma_{t-1} + \eta_{3t}, \end{cases}$$

where $\alpha_t,\ \beta_t,\ \gamma_t$ are interpreted as local intercept, slope and acceleration parameters respectively of the SSM, and η_{kt} $(k = 1,\ 2,\ 3)$ are uncorrelated random errors.

The prediction curve for some cancer sites displays excess variability. This is handled with a tuning parameter in the SSM prediction model. Technical details of incorporating the tuning parameter into the SSM can be found in Ghosh *et al.* (2007).

3.2 Validation method

The comparability of the death rate predictions from the two methods (PF and SSM) was assessed using a weighted average of the squared deviation differences between the one-, two- and three-year-ahead projected and observed values. Deviations for comparing different cancer types are weighted by the number of deaths for that cancer type in 2003 as given in *Cancer Facts and Figures.* Deviations across time are weighted by the number of deaths in a given year. In comparing the two average squared deviations, the one with a smaller value is an indicator of the predicted value falling closer to the observed value. When rates

are compared for both genders combined or for multiple cancers combined (since the death rates vary by cancer site and gender), weighted averages of the squared deviation differences were used, with the age-adjusted estimated death rate for the most recent year used as the weight.

4. Results

4.1 Model comparison and results

Figure 1 compares the predicted Proc Forecast (PF) and State Space Model (SSM) age-adjusted death rates for prostate cancer.

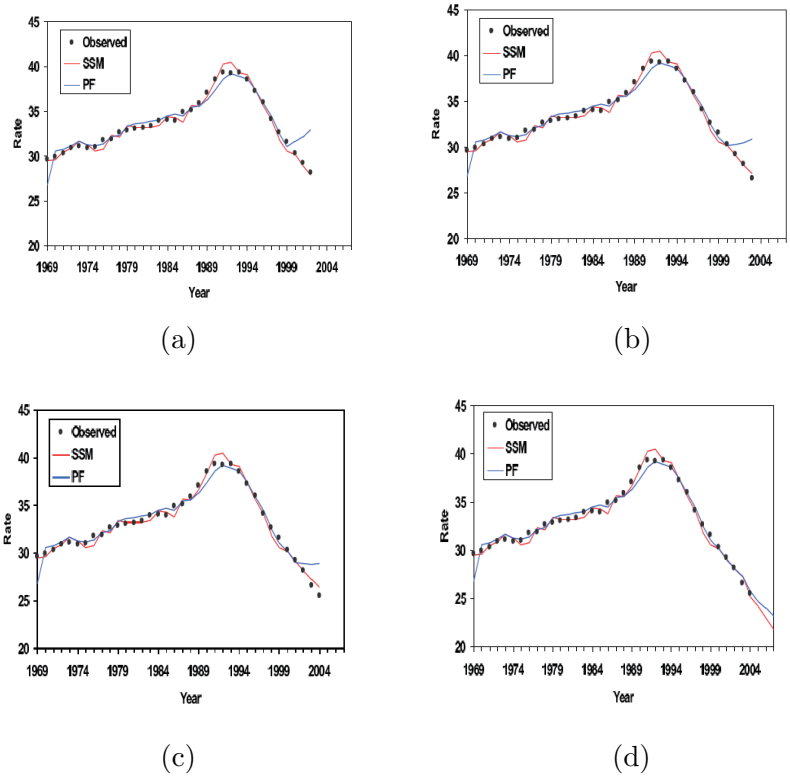


Figure 1: Age-adjusted death rate (per 100,000) predictions for prostate cancer: (a) fits for 1969-1999, projections for 2000-2002; (b) fits for 1969-2000, projections for 2001-2003; (c) fits for 1969-2000, projections for 2001-2003; (d) fits for 1969-2004, projections for 2005-2007.

In Figure 1a, we used the data on observed age-adjusted death rates for prostate cancers from 1969 to 1999 to fit the two models, and we extrapolated one-, two-, and three-year-ahead projections for 2000, 2001, and 2002. Both

models fit the observed data (1969 to 1999) fairly well. To be sure validation was spread across several calendar years, the analysis was repeated for subsequent years and is displayed in Figures 1a – 1d. In Figure 1b, we used observed data from 1969 to 2000 to extrapolate death rates for 2001, 2002, and 2003, and in Figure 1c, we used observed data from 1969 to 2001 to extrapolate death numbers for 2002, 2003, and 2004. Finally, Figure 1d shows how the actual projections would occur in practice, projecting out to the future where we currently have no data to validate the results. This panel uses the most recent available data (1969 to 2004) at the time this report was written and projects through 2007.

Figure 1a shows that the SSM method has strong ability to capture short term trends, whereas the PF predictions are poor and show an increasing trend while the observed values are actually decreasing. Figures 1b and 1c show extremely accurate predictions by the SSM method, with poor future predictions for the PF method.

Table 1: Observed and three-year-ahead predictions for cancer site / sex combinations for 2004, using data from 1969 to 2001.

	3-Yr Predicted Values		
Cancer Site/Sex Combinations	Observed	PF	SSM
Colon & Rectum (Females)	15.15	**15.56**	16.10
Lung & Bronchus (Females)	40.87	43.31	**41.33**
Melanoma of the Skin (Females)	1.70	1.69	1.69
Breast (Females)	24.38	**23.84**	23.80
Ovary (Females)	8.75	9.03	**8.82**
Hodgkin Lymphoma (Females)	0.34	0.49	**0.34**
Non-Hodgkin Lymphoma (Females)	5.68	7.15	**5.20**
Acute Lymphocytic Leukemia (Females)	0.39	0.45	**0.41**
Colon & Rectum (Males)	21.57	**21.21**	22.41
Lung & Bronchus (Males)	70.29	67.94	**70.57**
Melanoma of the Skin (Males)	3.94	**3.92**	3.83
Prostate (Males)	25.45	28.91	**26.41**
Testis (Males)	0.25	0.31	**0.23**
Hodgkin Lymphoma (Males)	0.54	0.69	**0.55**
Non-Hodgkin Lymphoma (Males)	8.85	11.13	**8.54**
Acute Lymphocytic Leukemia (Males)	0.55	0.64	**0.58**

Note: Bold text denotes the model fit with a result closest to the observed value (displayed results are rounded to the second decimal).

In order to compare the two methods proposed for projection of future age-adjusted cancer death rates, we modeled one-, two-, and three-year ahead projections for 2002 to 2004 based on data collected from 1969 to 2001. The observed

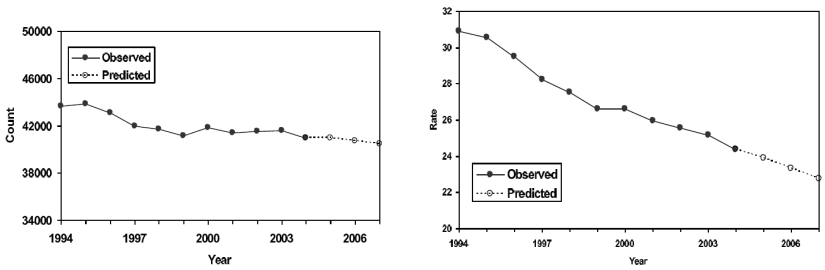
age-adjusted cancer death rates for 2002, 2003, and 2004 are available and are used to validate the prediction models. Table 1 displays the observed and predicted age-adjusted cancer death rates for the eight cancer sites with the highest number of deaths (American Cancer Society, 2007). For both females and males, the SSM method produced values closer to the observed age-adjusted cancer rates than the PF method for each cancer site. The SSM outperformed the PF method for all of the male types of cancer presented. One-, two-, and three-year-ahead predicted age-adjusted cancer rates for all cancer sites combined are displayed in Table 2.

Table 2: Observed and Three-year-ahead Predictions for all Sites Combined for Females and Males for 2002 to 2004.

	Observed	PF	SSM
Females 2002	162.87	168.43	**162.10**
Females 2003	160.45	166.78	**163.23**
Females 2004	156.96	164.19	**160.78**
Males 2002	240.20	244.52	**240.03**
Males 2003	234.12	238.74	**236.51**
Males 2004	228.26	233.11	**231.34**

Note: Bold text denotes the model fit with a result closest to the observed value (displayed results are rounded to the second decimal).

Observed and three-year-ahead predicted cancer death counts as well as death rates from the SSM and age-adjusted cancer death rates for breast cancer are shown in Figures 2.



(a) (b)

Figure 2: Observed data (1999-2004) and 1-, 2-, and 3-year-ahead predicted (2005-2007) count and rate values for breast cancer in females. (a) Female breast counts; (b) female breast rates.

The count and rate figures have appropriately scaled vertical axes so as to display proportional changes between the two measures. The number of breast

cancer deaths drops slightly over these years, although the population growth over these years is more substantial, giving a more rapidly declining rate than count. In contrast, the age-adjusted death rates show a steady decline over the same years. This illustration shows the importance for policy-makers and clinicians to consider both the number of deaths from cancer in the population and the death rates that consider the population size as it changes over time. Studying one without the other does not give a complete picture of the cancer death burden.

5. Discussion

This year, 559,650 Americans are expected to die of cancer (American Cancer Society, 2007). Policymakers need current age-adjusted rate estimates alongside the current age-adjusted count estimates, in order to make evidence-based policy decisions that consider population size and change. The cancer burden is growing, and based on age-adjusted incidence rates between 1998 and 2002, the number of cancer patients is expected to more than double from 1.36 million in 2000 to nearly 3.0 million in 2050 due to aging and the growing U.S. population (Hayat *et al.*, 2007).

National level age-adjusted death rates for the current year 2007 can be predicted based on observed data through 2004. We have described two statistical modeling approaches for accomplishing this. Comparison of the two methods was carried out using a weighted square deviation between the predicted and observed one-, two-, and three-year-ahead age-adjusted rates, and the results suggest the SSM model is performing better than the PF method. These results are in agreement with the modeling results comparing the PF and SSM methods for projecting national age-adjusted death counts (Tiwari *et al.*, 2004).

The SSM method presented here assumes the error variance in the measurement equation is constant. Assuming the counts are realizations of Poisson random variables, the error variances can be assumed to be time dependent, and given by

$$Var(r_t) = \sum_{j=1}^{J} w_j^2 \frac{d_{tj}}{n_{tj}^2}.$$

However, our analysis (details not presented here) showed that this model did not perform as well without the assumption of constant error variance. Furthermore, in order to implement the SSM method with non-constant error variance, knowledge of the denominator, n_{tj}, is needed for future years.

Acknowledgement

We thank Gretchen Keel from Information Management Systems, Inc. for

technical assistance. We also thank the Editor for several constructive comments which improved the presentation of the paper. Work of Ram C. Tiwari was conducted during prior employment at the National Cancer Institute. The views expressed are those of this author and do not necessarily reflect those of the US Food and Drug Administration.

References

American Cancer Society (2007). *Cancer Facts and Figures 2007*. American Cancer Society.

Ghosh, K. and Tiwari, R. C. (2007). Prediction of U.S. cancer mortality counts using semiparametric Bayesian techniques. *Journal of the American Statistical Association* **102**, 7-15.

Ghosh, K., Tiwari, R. C., Feuer, E. J., Cronin, K. A. and Jemal, A. (2007). Predicting US Cancer Mortality Using State Space Models. In *Computational Methods in Biomedical Research* (Edited by Naik, D. and Khattree, R), 131-151., Marcel Dekker/Francis & Taylor.

Hayat, M. J., Howlader, N., Reichman, M. E., and Edwards, B. K. (2007). Cancer Statistics, Trends, and multiple primary cancer analyses from the surveillance, epidemiology and end results (SEER) program. *Oncologist* **12**, 20-37.

Pickle, L. W., Feuer, E. J. and Edwards, B. K. (2003). *U.S. Predicted Cancer Incidence, 1999: Complete Maps by County and State from Spatial Projection Models*. no. 5 in NCI Cancer Surveillance Monograph Series, National Cancer Institute.

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

Wingo, P. A., Landis, S., Parker, S., Bolden, S. and Heath, G. W. (1998). Using cancer registry and vital statistics data to estimate the number of new cancer cases and deaths in the United States for the upcoming year. *Journal of Registry Management* **25**, 43-51.

World Health Organization (1975). *Manual of International Statistical Classification of Diseases, Injuries and Causes of Death*. World Health Organization, Geneva, Switzerland, vol. 1, 9th rev. ed.

World Health Organization (1992). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*. World Health Organization, Geneva, Switzerland, volume 1, 10th revision ed.

Received June 1, 2008; accepted November 1, 2008.

Matthew J. Hayat
School of Nursing
Johns Hopkins University
525 N. Wolfe St, Room 532
Baltimore, MD, 21205 USA
mhayat2@son.jhmi.edu

Ram Tiwari
Office of Biostatistics
Center for Drug Evaluation & Research
U.S. Food and Drug Administration
10903 New Hampshire Ave.,
WO Bldg. 21, Rm. 3524
Silver Spring, MD 20993, USA
ram.tiwari@fda.hhs.gov

Kaushik Ghosh
Department of Mathematical Sciences
University of Nevada Las Vegas
4505 Maryland Parkway
Box 454020, Las Vegas, NV 89154-4020, USA
kaushik.ghosh@unlv.edu

Mark Hachey
Information Management Services, Inc.
12501 Prosperity Dr, Suite 200
Silver Spring, MD 20904 USA
hacheym@imsweb.com

Benjamin F. Hankey
Information Management Services, Inc.
12501 Prosperity Dr, Suite 200
Silver Spring, MD 20904 USA
bhankey0411@aol.com

Eric J Feuer
Division of Cancer Prevention and Control and Population Sciences
National Cancer Institute
6130 Executive Boulevard, Room 6134
Rockville, Maryland 20852, USA
rf41u@nih.gov