CA

A Cancer Journal for Clinicians

A New Method of Estimating United States and State-level Cancer Incidence Counts for the Current Calendar Year

Linda W. Pickle, Yongping Hao, Ahmedin Jemal, Zhaohui Zou, Ram C. Tiwari, Elizabeth Ward, Mark Hachey, Holly L. Howe and Eric J. Feuer

CA Cancer J Clin 2007;57;30-42

CA: A Cancer Journal for Clinicians is published six times per year for the American Cancer Society by Lippincott Williams & Wilkins. A bimonthly publication, it has been published continuously since November 1950. *CA* is owned, published, and trademarked by the American Cancer Society, 1599 Clifton Road, NE, Atlanta, Georgia 30329. (©American Cancer Society, Inc.) All rights reserved. Print ISSN: 0007-9235. Online ISSN: 1542-4863.

A New Method of Estimating United States and State-level Cancer Incidence Counts for the Current Calendar Year

Linda W. Pickle, PhD; Yongping Hao, PhD; Ahmedin Jemal, DVM, PhD; Zhaohui Zou, MS; Ram C. Tiwari, PhD; Elizabeth Ward, PhD; Mark Hachey, MS; Holly L. Howe, PhD; Eric J. Feuer, PhD

ABSTRACT The American Cancer Society (ACS) has published the estimated number of new cancer cases and deaths in the current year for the United States that are commonly used by cancer control planners and the media. The methods used to produce these estimates have changed over the years as data (incidence) and statistical models improved. In this paper we present a new method that uses statistical models of cancer incidence that incorporate potential predictors of spatial and temporal variation of cancer occurrence and that account for delay in case reporting and then projects these estimated numbers of cases ahead 4 years using a piecewise linear (joinpoint) regression method. Based on evidence presented here that the new method produces more accurate estimates of the number of new cancer cases for years and areas for which data are available for comparison, the ACS has elected to use it to estimate the number of new cancer cases in *Cancer Facts & Figures 2007* and in *Cancer Statistics, 2007.* (*CA Cancer J Clin 2007;57:30–42.*) © American Cancer Society, Inc., 2007.

Dr. Pickle is Senior Mathematical Statistician and Coordinator of Geographic Research, Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD.

Dr. Hao is GIS Analyst and Statistician, Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA.

Dr. Jemal is Strategic Director, Cancer Occurrence, Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA.

Mr. Zou is Statistical Programmer, Information Management Services, Inc., Silver Spring, MD.

Dr. Tiwari is Mathematical Statistician and Program Director, Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD.

Dr. Ward is Managing Director, Surveillance Research, Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, GA.

Mr. Hachey is Statistical Programmer, Information Management Services, Inc., Silver Spring, MD.

Dr. Howe is Executive Director, North American Association of Central Cancer Registries, Inc., Springfield, IL.

Dr. Feuer is Chief, Statistical Research and Applications Branch, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD.

This article is available online at http://CAonline.AmCancerSoc.org

INTRODUCTION

The American Cancer Society (ACS) has published the estimated number of new cancer cases and deaths in the current year for the total United States and individual states in its annual *Cancer Facts & Figures* (CFF) publication since 1960 and in *Cancer Statistics* in this journal since the early 1970s.[1,2] These estimates for the current year provide important information for cancer control planners, public policy analysts, and state health departments, who need to know the relative burden of different types of cancer and for different geographic regions.

The method to produce the ACS estimates has been refined as more incidence data have become available and statistical methods have improved. Beginning with the 1998 estimates, the statistical projection methods for cancer cases and deaths were changed from linear projections to an autoregressive quadratic time trend model.[3] The projection method for deaths was further changed to a state-space model (SSM) beginning with the 2004 estimates, after a study demonstrated that the SSM produced more accurate predictions than the autoregressive quadratic time trend model.[4]

In order for the methods now used by the ACS to project accurate estimates of new cases and deaths to the current year, long-term data (8 or more years) must be available for all US states or for a subset of states that are representative of the entire United States. Long-term cancer mortality data exist for all US states since 1933, while long-term incidence data are available since 1975 only from the original registries included in the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program (SEER9), covering about 10% of the population.[5]

The ACS method projects the total number of cases in the United States to the current year by a two-step process. First, the annual age-specific rates in the 9 oldest SEER areas are applied to the corresponding age-specific population from 1979 to the most current year for which data are available to estimate the number of new cancer cases diagnosed in each of those years. Then, a quadratic autoregressive time series model is applied to these estimates to project 4 years ahead to produce the projected total number of cases in the current year. State estimates are derived by apportioning the total US case estimates by state, based on the distribution of estimated cancer deaths. Underlying assumptions of this method are that age-specific incidence rates from the combined 9 oldest SEER cancer registries are representative of the US population and that the incidence-to-mortality ratios are constant across all states.

Cancer registries have now been established in every state and territory in the United States, and high-quality incidence data are available for several years for most, providing the opportunity to improve the ACS case projections by taking geographic variability of incidence rates into account. However, since only about half of states outside the SEER9 areas have incidence data that have met national criteria of high quality and completeness for 8 or more years,[6] with no data available at all for some states, a new method for case projection was developed.

The new method uses statistical models of cancer incidence that incorporate potential predictors and spatial and temporal variation of cancer occurrence and that account for delay in case reporting. This paper describes the new method and compares its case projections for 2007 to those using the existing ACS method. Based on evidence that the new method produces more accurate estimates of the number of new cancer cases for years and areas for which data are available for comparison, the ACS has elected to use it to estimate the number of new cancer cases in CFF 2007 and in *Cancer Statistics, 2007*.[7,8]

MATERIALS AND METHODS

The detailed methodology used for the spatial component of the method has been presented elsewhere.[9] This method was previously used to produce state- and county-level maps of estimated cancer incidence in 1999[9]; a simpler ecologic regression model was shown to be a reasonable method for projecting breast cancer incidence rates from SEER areas to the entire United States for a single time period.[10] The number of cases for all US counties was estimated based on the association of demographic and lifestyle profiles with cancer occurrence in the 480 SEER counties (SEER17) in 1999 using a hierarchical Poisson regression model (L.W. P., unpublished data, 2006). This method was found to produce accurate estimates of state incidence counts and rates for all cancer sites combined, for the four most frequent cancers, and for all other cancer sites combined.

To validate the proposed methods for estimating the numbers of new cases in 2007, the spatial and temporal components of the method were tested separately. First, the spatial model described above was used to estimate the numbers of new cases in every US state for four major cancer sites (breast, prostate, lung and bronchus, colon and rectum) in each year for which state-specific results were available in the *U.S. Cancer Statistics Report* (USCS).[11–13] USCS reports included the numbers of cases for 25 types of cancer reported by 42 states in 1999 and 2000 and by 44 states in 2001. This test was based on the 17 SEER registries with data available for each test year. Output from this model consisted of the numbers of cases estimated for each state that year; these are either modeled estimates for states that have data or "spatial projections," ie, estimates for states that have no observed data for a given year, based on data available from other registries. For comparison, the numbers of cases were also estimated for each state and year using the previous ACS method. Results from each method were compared with the observed numbers of cases as published in the USCS reports either by the squared deviations (square of the estimated minus observed counts) of the total summed over available states or by the sum of the squared deviations for each state.

As a second step in the validation process, output from the spatial projection model applied to each of a number of years was used to find which temporal projection method was best for projecting incidence counts 4 years ahead in time. This study was based on observed numbers of malignant cases from the SEER registries beginning in 1988 (with varying numbers of registries over time as SEER expanded from SEER9 to SEER17).[5] Data from 1988 to 1995 were used to predict the 1999 estimated number of new cases, from 1988 to 1996 to predict 2000, and from 1988 to 1997 to predict 2001.

Four different methods for temporal projection of model-based estimates were tested: the previous ACS quadratic time series method (PROC FORECAST [PF]), a state-space method (SSM) currently used to project mortality counts ahead in time for CFF,[4] a piecewise linear regression method (joinpoint method [JP])[14,15] currently used to describe trends in incidence and mortality in many cancer registry reports,[16] and a newly proposed semiparametric Dirichlet process method (DIR).[17] Each of these methods was used to determine the time trends in the estimated counts across the available data years, then to project the number of cases 4 years ahead. The projected state-specific numbers of cases from each method were compared with the observed numbers of cases as published in the USCS reports on the basis of the sums of squared deviations.

Following the determination of the best spatial models and temporal projection method, the model was extended to incorporate time trends over the data period (L.W. P., unpublished data, 2006). The time trend was modeled as a quadratic function, similar to the previous ACS method, but the temporal effect could vary by geographic region or by county characteristic (eg, time trends could differ in urban and rural counties). The model included extra variation due to correlation of the numbers of cases over time and place (county, state, and region) and an additional term to account for any remaining "overdispersion," ie, greater than expected variation in Poisson-distributed counts. This model was implemented using SAS PROC GLIMMIX software with its optional spline-based approximation for spatial and temporal autocorrelation[18] (also L.W. P., O. Schabenberger, A. Stephens, unpublished data, 2006). One advantage of this more complex spatio-temporal model is that only a single application of the model to data for the entire time span is required, rather than separate applications of the model to each year's data. More importantly, the spatio-temporal model shares information across nearby points of time and place simultaneously to provide the best results.

The spatial projection component of the model, ie, estimation of numbers of new cases in states without observed data, requires good spatial coverage in all regions of the United States, so utilizing data from a large and geographically dispersed portion of the United States was critical. For the 2007 projection, an incidence database covering 1995 to 2003 was obtained through an agreement with the North American Association of Central Cancer Registries (NAACCR). The data source was the response to the NAACCR Call for Data submissions as of December 2005. US cancer registries reporting data to NAACCR participate in the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program or the Centers for Disease Control and Prevention (CDC) National Program of Cancer Registries (NPCR), or both, and receive support from the state, province, or territory where they are located. Registries for 40 states, the District of Columbia (DC), and the Detroit metropolitan area (Figure 1) met NAACCR registry certification standards as providing complete, accurate, and timely data for at least 3 consecutive years during 1995 to 2003[6] and agreed to release county-level incidence data for this project. Together, these registries cover 86% of the US population, although not every state included in this modeling effort had data for every year.

Projections were made initially at the county level to capture within-state variation of incidence counts. However, for 19 of the rarest sites included in CFF at the national level, projections were made at the Health Service Area (HSA) level to provide more stable counts for analysis. The HSA units are groups of counties defined on the basis of where county residents aged 65 years and over obtained hospital care, and have been used in other geographic analyses.[19,20] Case records without a valid county identifier were randomly assigned to a county in proportion to the distribution of cancer cases with known county of residence in that state.

The cancer site was coded according to the SEER Program recodes in the same manner used

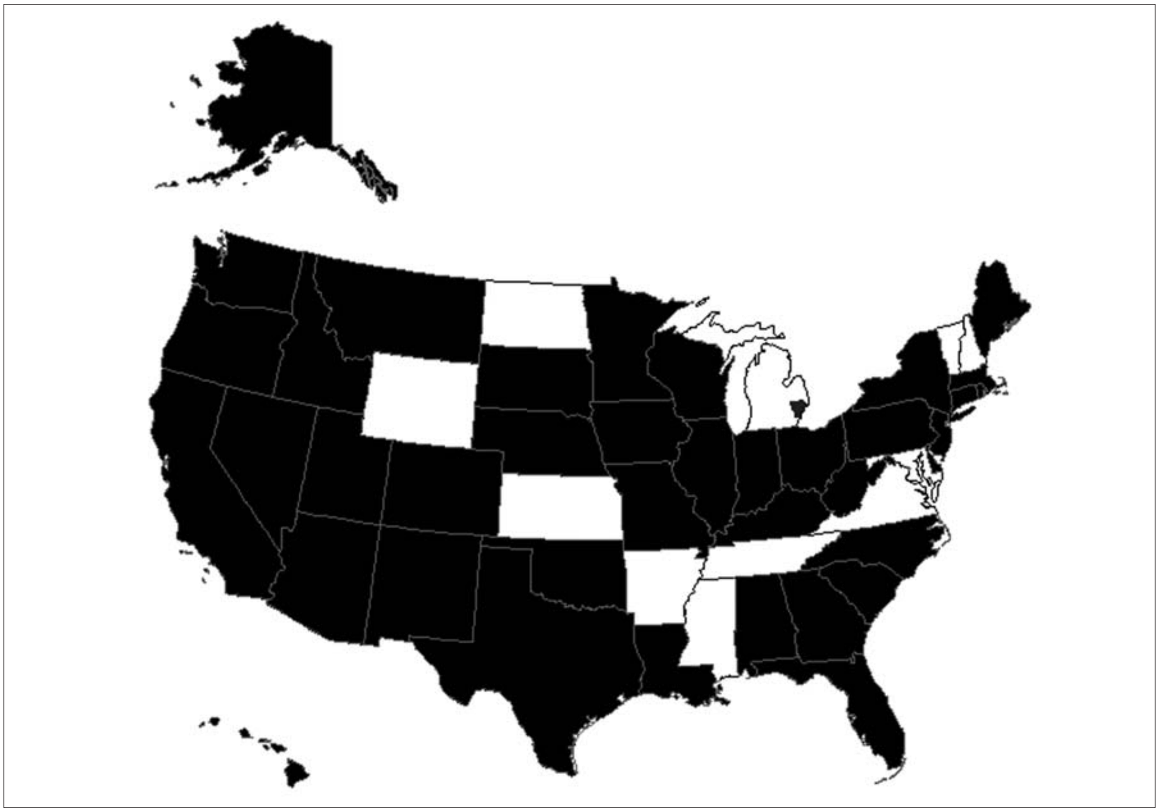

FIGURE **1** Shaded Areas Indicate Registries Included in 1995 to 2003 Cancer Incidence Database Available From the North American Association of Central Cancer Registries (NAACCR).

for previous CFF reports.[21] Race was grouped as White, Black, and Other. Although the NAACCR file identifies much finer race categories, the numbers of cases observed among Hispanics and Asian American/Pacific Islanders, for example, were too low in most regions of the United States to permit stratification of individual cases beyond three broad categories. However, the percentages of Hispanics, Asian American/Pacific Islanders, and American Indian/Alaskan Natives in each county were included in the model to capture variations in incidence due to different racial mixes of the population. Age at diagnosis was initially coded to age groups 0 to 4 years, 5 to 14 years, 15 to 24 years, 25 to 34 years, 35 to 44 years, 45 to 54 years, 55 to 64 years, 65 to 74 years, 75 to 84 years, and 85+ years; younger age groups were usually aggregated to ensure adequate numbers of cases in each stratum for analysis, typically age 0 to 34 years, depending on the cancer site. Input to the models consisted of numbers of new cases stratified by site, sex, race, age group, county or HSA of residence, and year of diagnosis rather than individual case records. Similarly stratified populations were obtained from the Census Bureau.[22]

Approximately 35 covariates were considered as potential predictors of incidence in the new models. Only age, sex, race, county of residence, and type of cancer were available for the individual cases. All other predictors were population characteristics for the county or HSA, including measures of income, education, housing, racial distribution, urban/rural status, availability of physicians and cancer screening facilities, health insurance coverage, cigarette smoking, obesity, cancer screening rates, and mortality rates. These covariates were available for every US county from a variety of sources, including the Census Bureau, Area Resource File,[23] CDC,[24] and the National Center for Health Statistics.[25] Behavioral risk factor and screening variables from the CDC Behavioral Risk Factor Surveillance System were calculated as mean proportions at the state level for each year. Differences between each county's calculated proportion and its state value for the aggregated period 1994 to 2003 were also calculated to measure within-state variation of the risk and cancer screening behaviors. Annual values for all other covariates were calculated by linear interpolation between available data years and linear extrapolation to 2003 beyond the last available year.

Results of the spatio-temporal models are cancer- and sex-specific smoothed annual estimates for registries that provided data and annual modeled estimates for registries with missing data for each year (1995 to 2003). The assumed spatial and temporal autocorrelation plus covariates included in the model result in a sharing of information across areas that are similar in location, time, and county characteristics. For example, the number of new cases for registries with no input data at all will be estimated using several years of data from neighboring states and from other states and counties with similar sociodemographic and lifestyle profiles; estimated numbers for a registry with a single missing year of data are based on observations from that registry before and after the missing time point, as well as from states that are neighbors or have similar characteristics.

Model estimates were added over age, race, and county to produce state-year-cancer-specific estimates for the time span of the available incidence data. These estimated numbers were then adjusted to account for the delay expected in reporting cancer cases to the registry.[26] The number of new cases reported to the SEER registries in the most recent data year are on average 3.5% to 4.5% below what they eventually will be after case finding by the registry is complete, but can range as high as 21% (for leukemia), depending on the type of cancer and the sex, race, and age of the patient. The delay adjustment modifies the observed numbers more in the most recent reporting years to account for future anticipated corrections to the data.[27] To date, delay adjustment estimates have only been developed for the long-running SEER9 registries. However, results from all registries, not just SEER9, were delay adjusted, assuming that these SEER-derived factors hold for the entire United States. As longer incidence time series are available from more registries, more appropriate delay factors can be developed. Although the factors used in this new method are not ideal, without any adjustment at all the number of new cases could falsely appear to be trending downward in the most recent years, impacting the projected trend into the future. The delay-adjusted numbers were then projected ahead to 2007.

RESULTS

Table 1 and Table 2 present the results of the validation study of the spatial and temporal components of the model, respectively. Table 1 presents the numbers of new cases reported for the 4 most common cancer sites by sex for the sum of all states in USCS for each year from 1999 to 2001, as estimated by the previous ACS method and the new spatial projection model. The spatial model estimates for female breast and male lung cancers were closest to the observed values for all 3 estimated years (1999 to 2001), whereas for prostate cancer, the estimates from the ACS method were the closest to the observed values. Estimates for colon cancer were equally discrepant from the observed values, with the ACS method underestimating and the new method overestimating the reported numbers. Overall, estimates based on the new method were closest to the observed values for 8 of the 18 estimates. However, for state-specific estimates, the new method provided closer estimates to the observed values than the old method, occasionally by a wide margin. For example, the sum of state-specific squared deviations averaged over 1999 to 2001 for female breast cancer was 20 times higher in the ACS method compared with the new method (Table 1).

For the temporal method validation, non-Hodgkin lymphoma, melanoma, esophageal cancer, and testicular cancer were added to test performance for rarer sites. The spatial projection model was used to estimate the numbers of new cancer cases for each year, beginning in 1988, and these estimates were then used to project the numbers 4 years ahead using 4 temporal projection methods. As shown in Table 2, the JP provided projections closest to the observed total number of new cases for 7 of the 13 comparisons, DIR was best for 4, and PF for 2. Similar results were found for the other 2 years the projections were done (1999 and 2000). The JP was by far the best projection method when state-specific squared errors were compared.

TABLE **1** Validation of Spatial Model: Observed and Predicted Numbers of New Cases for 1999 to 2001 by Cancer Site and Sex, With Sums of State-specific Squared Deviations From Observed Number of New Cases for Registries Reporting in *U.S. Cancer Statistics Report* (USCS), Averaged Over 1999, 2000, and 2001

Sex/Cancer Site/Year	**USCS**	**Estimated Number of Cases Summed Over USCS Registries**		**Sum of State-specific Squared Deviations, Averaged Over 1999 to 2001**	
	Observed	**ACS Method**	**Spatial Model**	**ACS Method**	**Spatial Model**
Female					
Breast				36,774,000	1,817,554*
1999	174,912	178,478	176,045*		
2000	164,895	166,989	165,721*		
2001	180,038	184,321	178,833*		
Lung and bronchus				5,371,894*	8,333,344
1999	72,999	68,120*	80,369		
2000	69,316	64,064*	77,550		
2001	77,221	68,939	82,624*		
Colon and rectum				3,306,224*	3,573,888
1999	67,186	64,692*	70,159		
2000	63,907	60,876*	66,977		
2001	68,513	65,237*	72,932		
Male					
Prostate				49,125,857	21,639,454*
1999	173,263	181,768*	183,717		
2000	164,753	177,619*	179,034		
2001	183,643	194,975*	195,966		
Lung and bronchus				12,702,367	11,499,802*
1999	96,468	82,596	100,557*		
2000	89,489	77,397	98,826*		
2001	98,687	83,728	105,589*		
Colon and rectum				3,306,695*	3,443,079
1999	68,166	64,868*	73,074		
2000	64,967	60,967	67,537*		
2001	69,822	65,314*	74,432		

*Indicates best results: estimated number of cases closest to observed values or lowest average sum of squared deviations.

Source of input to models is incidence data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program registries 1988–2001.[5]

The JP is more flexible than the PF because it fits multiple linear segments to the time series, and thus is more sensitive to sudden changes in trend than the presumed quadratic time trend used by PF. The semiparametric method (DIR) and the SSM apparently require a longer time series than was available in order to project several years ahead and cannot provide state-specific estimates for missing data states. On the basis of this validation study, the JP is the preferred method to project the number of new cases ahead in time, at least until a much longer time series is available for most states. Therefore, the projected numbers of cancer cases in 2007 for each sex/cancer site combination were produced by the following steps:

(1) apply the spatio-temporal model to data from registries available from NAACCR for 1995 to 2003 to estimate the numbers of cases by sex, race, age group, and year for every US county;

(2) sum these estimates over race, age, and county to calculate estimated numbers by sex and year for every US state and for the total United States, then adjust these results to account for case reporting delay;

(3) apply the JP to extrapolate each state's or US time series of the delay-adjusted estimated number of new cases to 2007.

Table 3 presents the results projected for the total United States by cancer site using the new and old methods for 2007. The new method

TABLE **2** Validation of Temporal Projection Method: Observed and Predicted Numbers of New Cases for 2001 by Cancer Site and Sex, With Sums of State-specific Squared Deviations From Observed Number of New Cases for Registries Reporting in *U.S. Cancer Statistics Report* (USCS), Averaged Over 1999, 2000, and 2001

	2001 USCS	**2001 Estimated Number of Cases Summed Over USCS Registries**				**Sum of State-specific Squared Deviations, Averaged Over 1999 to 2001**			
Cancer Site	**Observed**	**SSM**	**JP**	**PF**	**DIR**	**SSM**	**JP**	**PF**	**DIR**
Female									
Breast	180,038	188,673	181,154*	189,430	186,987	62,039,098	7,515,979*	11,812,999	9,426,462
Lung and bronchus	77,221	69,011	89,945	44,531	80,389*	137,883,200	22,217,849	541,501,075	17,540,282*
Colon and rectum	68,513	83,447	72,586	70,468*	76,833	31,986,025	2,541,926*	9,061,901	7,334,034
Non-Hodgkin lymphoma	22,399	28,617	23,858*	27,915	27,888	8,839,760	150,428*	1,638,051	1,833,078
Melanoma	18,019	22,461	16,865*	21,793	20,314	7,890,147	479,010*	2,055,440	1,562,063
Esophagus	2,951	2,637	3,142	3,682	2,816*	270,614	23,378*	145,671	98,089
Male									
Prostate	183,643	223,541	162,521	39,000	192,557*	1,334,558,296	365,082,240	1,648,765,768	117,864,930*
Lung and bronchus	98,687	132,850	98,854*	94,658	104,606	1,204,137,074	6,415,843*	185,851,928	84,090,958
Colon and rectum	69,822	87,012	77,278*	82,779	83,933	40,230,943	9,281,194*	18,002,668	14,532,673
Non-Hodgkin lymphoma	24,978	23,086	25,133*	23,723	24,428	6,097,704	718,596*	1,579,681	1,041,678
Melanoma	26,325	28,898	24,379	22,991	27,600*	54,544,440	17,315,126*	23,280,777	19,094,648
Esophagus	9,585	10,811	9,604*	10,339	9,751	4,313,903	71,510*	771,490	565,223
Testis	6,912	3,303	9,307	7,415*	5,351	20,256,415	439,004*	1,761,048	646,233

SSM = state-space model.
JP = joinpoint method.
PF = PROC FORECAST; previous ACS quadratic time series method.
DIR = semiparametric Dirichlet process method.
*Indicates best results: estimated number of cases closest to observed values or lowest average sum of squared deviations.
Source of input to models is incidence data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program registries 1988–2001.[5]

projects that there will be 1,444,913 new cancer cases among men and women in 2007, which is 1.8% higher than the 2007 projection using the previous ACS method. The total number of cases estimated by the old and new methods are quite similar over the period of 1995 to 2003, although the faster increase over time estimated by the new method leads to the slightly higher projected number of cases in 2007 (Figure 2). However, there are substantial differences between the two methods in the number of cases projected by site. Among the 4 most common cancer sites, projections from the new method compared with the old method are 15.3% higher for lung cancer, 3.7% higher for colorectal cancer, 5.5% lower for prostate cancer, and 15.2% lower for female breast cancer (Table 2). Cancer site groupings where the estimates of new cases are more than 10% higher than the previous CFF method predicted are oral cavity and pharynx (+11.1%), with a notable increase in pharyngeal cancer (+29.6%); respiratory system cancers (+17.1%), with notable increases in all 3 cancer site categories; urinary system (+11.7%), with a notable increase in cancer of the kidney and renal pelvis (+23.1%); multiple myeloma (+18.4%); and leukemia (+23.4%), with notable excesses in all 4 major subtypes of leukemia. There were 2 cancer site groupings where the estimates of new cases are more than 10% lower than the previous CFF method predicted; these are bones and joints (–11.1%) and female breast cancer (–15.2%), as previously noted. Differences of more than 10% lower were also observed for relatively uncommon cancer sites (Table 3).

In an effort to understand the reasons for these differences, estimates from the previous and new methods were plotted over the available time span. Using breast cancer as an example, Figure 3 shows that the estimated numbers of cases were fairly close during 1995 to 2003, but that the projected trend from 2003 to 2007 differed between the methods. Separating the data by registry group indicates that there were lower female breast cancer incidence rates in the other cancer registries in the United States compared with the rates in the nine oldest SEER registries (Figure 4). However, during 1999 to 2003, when 40 of the 41 states had observed numbers of cases for comparison, estimates from the new

TABLE **3** Comparison of Predicted Number of New Cancer Cases by Method for All Sites, Both Sexes, in 2007

	New Method	Previous ACS Method	Percent (%) Difference
All sites	1,444,913	1,419,000	1.8
Oral cavity and pharynx	34,346	30,920	11.1
Tongue	9,798	9,410	4.1
Mouth	10,652	9,930	7.3
Pharynx	11,798	9,100	29.6
Other oral cavity	2,099	2,480	–15.4
Digestive system	271,254	265,230	2.3
Esophagus	15,558	14,170	9.8
Stomach	21,259	21,960	–3.2
Small intestine	5,637	6,610	–14.7
Colon	112,347	108,000	4.0
Rectum	41,418	40,210	3.0
Anus, anal canal, and anorectum	4,653	4,840	–3.9
Liver and intrahepatic bile duct	19,155	20,800	–7.9
Gallbladder and other biliary	9,251	8,470	9.2
Pancreas	37,176	34,130	8.9
Other digestive organs	4,800	6,040	–20.5
Respiratory system	229,393	195,840	17.1
Larynx	11,295	8,330	35.6
Lung and bronchus	213,374	184,990	15.3
Other respiratory organs	4,724	2,520	87.5
Bones and joints	2,373	2,670	–11.1
Soft tissue (including heart)	9,220	9,550	–3.5
Skin (excluding basal and squamous)	65,062	68,420	–4.9
Melanoma-skin	59,944	61,850	–3.1
Other nonepithelial skin	5,118	6,570	–22.1
Breast	180,506	212,950	–15.2
Genital system	306,388	318,940	–3.9
Uterine cervix	11,152	10,100	10.4
Uterine corpus	39,079	40,580	–3.7
Ovary	22,433	20,550	9.2
Vulva	3,494	4,220	–17.2
Vagina and other genital, female	2,145	2,410	–11.0
Prostate	218,885	231,550	–5.5
Testis	7,917	8,050	–1.7
Penis and other genital, male	1,284	1,480	–13.2
Urinary system	120,396	107,820	11.7
Urinary bladder	67,154	63,710	5.4
Kidney and renal pelvis	51,190	41,590	23.1
Ureter and other urinary organs	2,053	2,520	–18.5
Eye and orbit	2,336	2,490	–6.2
Brain and other nervous system	20,496	19,570	4.7
Endocrine system	35,521	35,020	1.4
Thyroid	33,555	32,610	2.9
Other endocrine	1,966	2,410	–18.4
Lymphoma	71,371	68,240	4.6
Hodgkin disease	8,189	7,600	7.7
Non-Hodgkin lymphoma	63,182	60,640	4.2
Multiple myeloma	19,907	16,810	18.4
Leukemia	44,236	35,840	23.4
Acute lymphocytic leukemia	5,206	4,070	27.9
Chronic lymphocytic leukemia	15,335	11,640	31.7
Acute myeloid leukemia	13,404	11,560	16.0
Chronic myeloid leukemia	4,572	4,050	12.9
Other leukemia	5,719	4,520	26.5
Other and unspecified primary sites	32,108	28,690	11.9

Results were produced by the previous *Cancer Facts & Figures* (CFF) method ("previous ACS method") and the new method of spatio-temporal model estimates, plus projection ahead in time by the joinpoint method (JP) ("new method"). Note the figures presented in Table 3 do not exactly equal those published in *Cancer Statistics, 2007* or CFF, where the numbers of new cases are rounded to the nearest 10.

Source of input to models is incidence data from the North American Association of Central Cancer Registries (NAACCR) Call for Data submissions as of December 2005.[6]

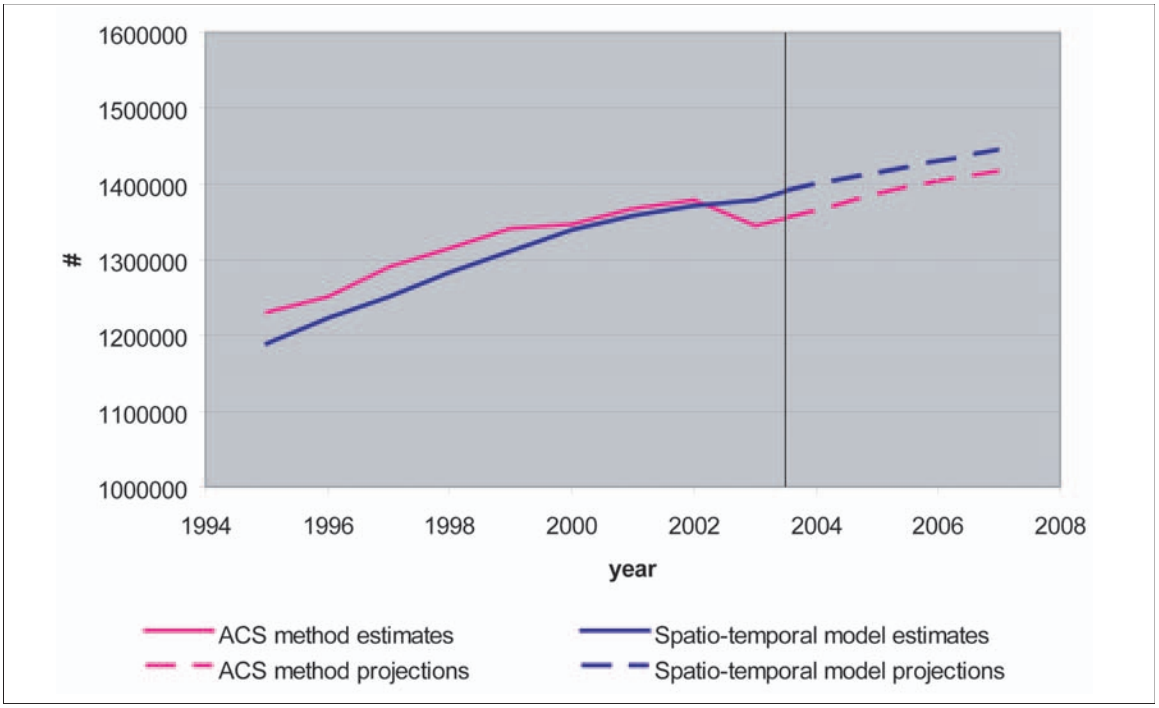

FIGURE 2 Comparison of Estimates and Projections of the Total Number of New Cancer Cases for All Sites, for Males and Females Combined by Year, Produced by the Previous *Cancer Facts & Figures* (CFF) Method and the New Method of Spatio-temporal Model Estimates Plus Projection Ahead in Time by the Joinpoint Method (JP). Vertical line is shown at midpoint between the 2003 estimate and the 2004 projection. Source of input to models is incidence data from the North American Association of Central Cancer Registries (NAACCR) Call for Data submissions as of December 2005.[6]

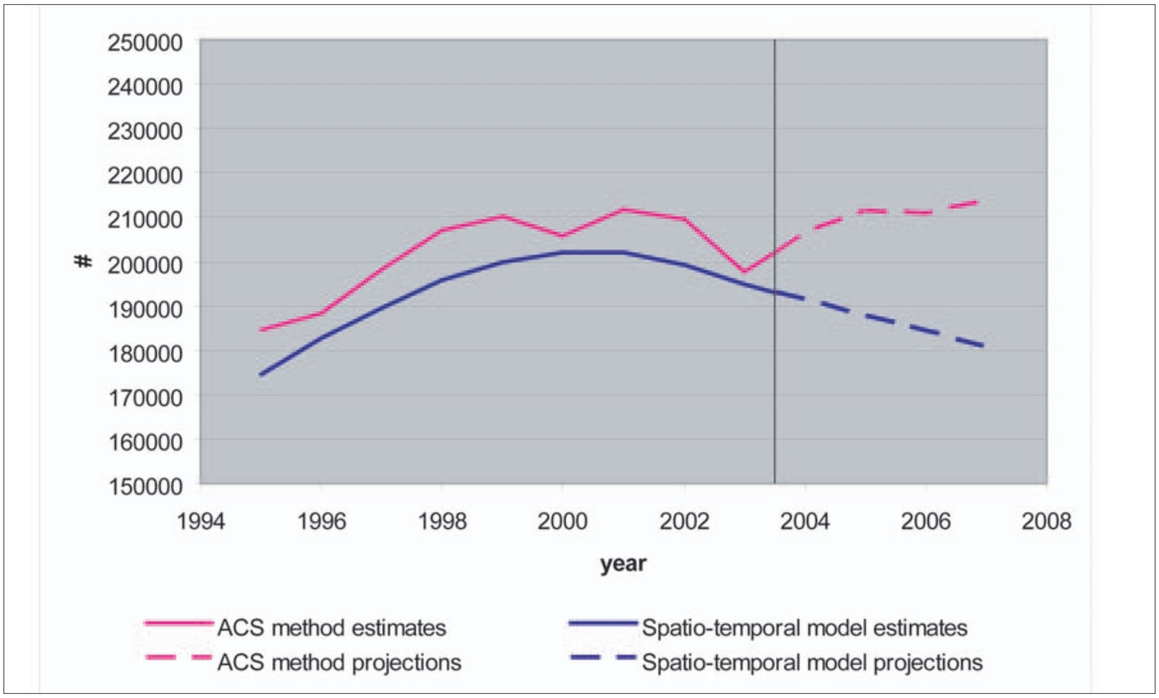

FIGURE 3 Comparison of Estimates and Projections of the Total Number of New Cases of Female Breast Cancer by Year Produced by the Previous *Cancer Facts & Figures* (CFF) Method and the New Method of Spatio-temporal Model Estimates Plus Projection Ahead in Time by the Joinpoint Method (JP). Vertical line is shown at midpoint between the 2003 estimate and the 2004 projection. Source of input to models is incidence data from the North American Association of Central Cancer Registries (NAACCR) Call for Data submissions as of December 2005.[6]

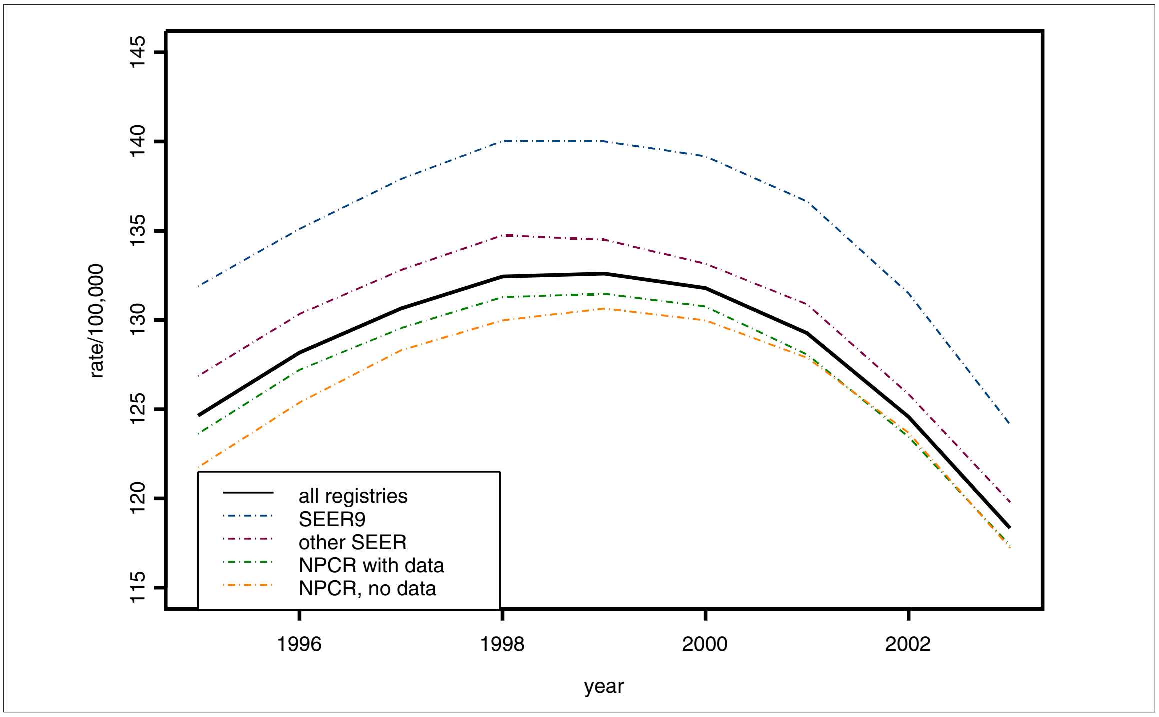

FIGURE **4** Estimated Age-adjusted Rates of Female Breast Cancer Incidence by Year Produced by the Spatio-temporal Model, According to Source of Data and Whether Data Were Available From the Registry. Source of input to models is incidence data from the North American Association of Central Cancer Registries (NAACCR) Call for Data submissions as of December 2005.[6]

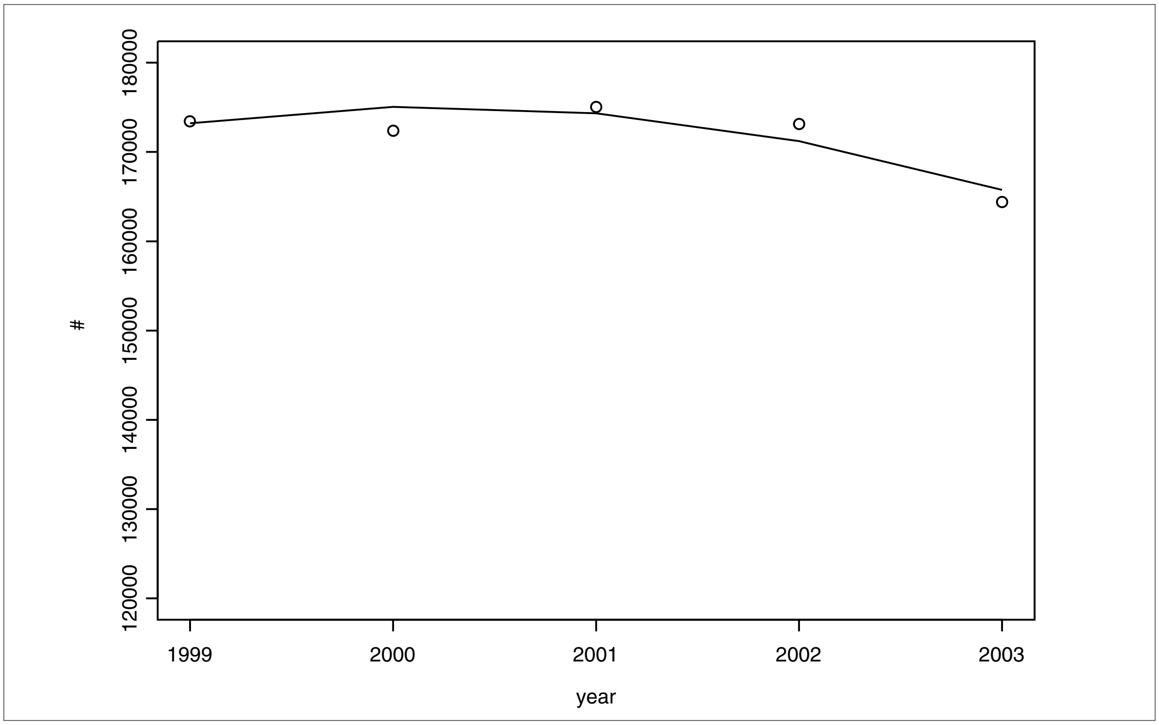

FIGURE **5** Numbers of New Cases of Breast Cancer Among Women Observed in Registries With Data Included in the NAACCR 1995–2003 File (Circles) Compared With Estimates Produced by the Spatio-temporal Model (Line). Estimates are not adjusted for potential delay in reporting. Data are only shown for 1999 to 2003 when nearly all registries provided data. Source of input to models is incidence data from the North American Association of Central Cancer Registries (NAACCR) Call for Data submissions as of December 2005.[6]

spatio-temporal model matched the observed data well (Figure 5).

Plots of observed and estimated numbers of cases for other types of cancer also suggested that the new model fit the observed data well (data not shown). Tables that compare the number of cancer cases projected by state by the old and the new method are available on the Web site (http://www.cancer.org/docroot/stt/stt_0.asp), along with a comparison of model-projected and observed cancer cases for 2003, the most recent year for which observed data are available.

DISCUSSION

The new spatio-temporal method with JP projection 4 years ahead shows a slightly greater increase in the total numbers of cases from 2006 to 2007 than would have been observed using the previous ACS method. The estimated number of cancer cases in 2006 was 1,399,790; the 2007 estimate using the old method is 1,419,000, a 1.4% increase. Thus, the 3.2% increase between the estimates of total new cases published in *Cancer Statistics*[2] and CFF[1] in 2006 and the estimate based on the new method in 2007 (1,444,913) reflects in part the discontinuity resulting from introduction of the new method and in part the underlying continuing increase in estimated cases resulting from growth and aging of the population.

Projections for specific cancer sites vary more substantially than projections of total cases. There are several reasons why the projections from the new method are likely to be more accurate than those from the earlier method:

- The new model allows for geographic variation in the underlying cancer incidence rates, consistent with observed incidence and mortality patterns, while the previous ACS method assumes that the age-specific incidence rates from the combined nine oldest SEER cancer registries are representative of the US population. Substantial geographic variability in cancer incidences rates has been demonstrated for many cancer sites.[13,28]
- The observed incidence data on which the 2007 estimates are based cover 86% of the US population, compared with 10% coverage by SEER9 registries. These expanded data are more representative of the United States, eg, the aggregated SEER9 registries were more urban and affluent and had lower proportions of tobacco users than the United States on the whole, so that cancers related to these factors might not be well estimated by the previous ACS method. Another consequence of greater geographic coverage is that fewer missing states need to be estimated.
- The previous ACS method assumes that the incidence-to-mortality ratios are constant across all states. However, mortality is poorly linked to incidence for less fatal cancers, and survival rates for most cancers have been demonstrated to vary geographically,[29] depending on screening patterns, the stage of disease at diagnosis, availability of treatment options, affordability of cancer care, comorbidity, socioeconomic factors, etc.
- The new spatio-temporal model includes many sociodemographic, medical facility, lifestyle, and cancer screening behavior variables as predictors of incidence; these factors were not considered in the previous ACS method.
- The new model includes spatial and temporal autocorrelation, ie, the method "borrows information" from other observations close in time and/or space. Thus, it can fill in "holes" in a state's time series, eg, before its becoming a certified high quality registry, or fill in "holes" in the map for a year when some states did not report their number of new cases. It has been shown that when autocorrelation is present, it must be accounted for in the model or results can be incorrect.[30]
- The new method inflates the reported number of cases to account for expected delays in case reporting. Thus, these new figures should better reflect the actual number of cases in the US population, once case finding is complete.

Like any method for projecting the number of new cancer cases 4 years ahead from observed data, the new method also has some limitations. Not all states and cancer sites are predicted equally well. The accuracy of the model results is dependent on inclusion of a sufficient set of covariates to explain the incidence patterns across the United States. The numbers of new cancer cases can be adequately predicted for most states using the new model, even without observations from them, but the presence of unmeasured risk factors or effective cancer control programs can impact the number of cases in ways that cannot be predicted. For example, a model using data from NAACCR 1995 to 2002, which did not include data from Pennsylvania, substantially underestimated the number of new lung cancer cases among males in Pennsylvania, but when Pennsylvania data were included in the expanded dataset used for the 2007 projections, its predicted count was very close to the observed count for 2003. North Carolina, on the other hand, was well estimated whether or not its observed data were included as input to the model.

Another limitation of the new model as implemented for 2007 is its assumption of a quadratic time trend over the short time span of data (1995 to 2003). Although no evidence was seen for a lack of fit, this assumption may impose a curvature onto the time trend that is not present in the observed data and which limits the sensitivity of the model to short-term variations or sudden changes in the trend. In the future, as the time span of the data available from most state registries lengthens, improved time series models can be used.

Inaccurate projections of the numbers of cases to 2007 may result from applying delay-adjustment factors that are based on case finding patterns in SEER registries to all registry data. When additional information on cumulative reporting patterns is available for other areas, more appropriate factors can be used.

Large differences in projections by the old and the new methods for the major cancer sites are of special importance since they have the greatest impact on the cancer burden. The 15.3% increase in estimates of lung cancer cases in the new compared with the old method most likely results from recognized differences in tobacco use patterns between the SEER9 areas and the fuller geographic data set used in the new model. Average annual age-standardized lung cancer incidence rates (1999 to 2003) for males and for females in the 41 states providing input to the new method are 11% (male) and 5% (female) higher than those in the 9 oldest SEER areas used by the old method. Several other smoking-related cancers showed similar patterns.

The greater number of cases projected for leukemia and all of its subtypes appears to be due to the effect of delay adjustment, which was not included in the previous ACS method. Before projection to 2007, model estimates of the number of leukemia cases in 2003 were inflated by 10% for cases under age 45 years, by 21% for age 45 to 64 years, and by 18% for cases over age 64 years, resulting in a 12% greater total number of leukemia cases estimated in 2003 and 23% greater in 2007 by the new method. These factors have been used for several years to adjust SEER incidence rates that, for leukemia, can result in an apparent increasing trend when the observed rate trend is declining.[28] The long estimated delay in case reporting is due to the nature of cancers of the hematopoietic system. Because no surgery is required for diagnosis or treatment of leukemia, many cases are not seen in a hospital, making case finding more difficult for the cancer registry. Also, children and young adults are diagnosed with acute more often than chronic leukemia. These younger cases often initially present with a medical crisis and so are identified by a hospital record more often than older cases with chronic disease. Because of the new adjustment for these expected delays in case finding, the number of cases projected by the new method should better reflect the actual number of new leukemia cases.

For breast cancer, the reasons for the 15.2% decrease in projected cases for 2007 using the new compared with the old method may be somewhat more complex. Age-adjusted rates in SEER9 registries, which are the basis for the previous ACS method, were about 6% higher than similarly adjusted rates in the geographic areas used for the spatio-temporal model (Figure 4), suggesting that use of an expanded registry database is at least partly responsible for the lower projected number of breast cancer cases. Another factor that may contribute to the differences is the uncertainty in projecting ahead in time when the underlying incidence trends appear to be changing. Trends in breast cancer incidence rates in most geographic areas used as input to the spatio-temporal model have shown a recent stabilization, possibly even a downturn, after increasing for several years.[16] These changes have been modeled differently by the methods used to project numbers of cases to 2007 (Figure 3), and at the present time it is unclear which method is more accurate. However, for 1999 to

2003, the observed numbers of new cases in the geographic areas whose incidence data were used in the spatio-temporal model were well fit by the new model (Figure 5).

The lower case estimate (by 5.5%) for prostate cancer by the new method is in part due to regional differences in prostate cancer incidence rates covered by the 2 methods. Average annual age-standardized prostate cancer incidence rates for 1999 to 2003 in the 41 states providing input to the new method are 8.8% lower than that of the 9 oldest SEER areas used by the old method, perhaps reflecting regional differences in utilization of prostate specific antigen.

Despite some limitations, the new spatio-temporal model plus JP regression for temporal extrapolation appears to provide improved estimates of the numbers of new cases, both for individual states and for the nation, even for the less common cancers. Based on these results, the ACS has decided to use this method to project incidence number of new cases for CFF 2007.

REFERENCES

1. American Cancer Society. Cancer Facts & Figures 2006. Atlanta, GA: American Cancer Society; 2006.

2. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2006. CA Cancer J Clin 2006;56:106–130.

3. Wingo PA, Landis S, Parker S, et al. Using cancer registry and vital statistics data to estimate the number of new cancer cases and deaths in the United States for upcoming year. J Reg Management 1998;25:43–51.

4. Tiwari RC, Ghosh K, Jemal A, et al. A new method of predicting US and state-level cancer mortality counts for the current calendar year. CA Cancer J Clin 2004;54:30–40.

5. National Cancer Institute, U.S. National Institutes of Health. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence-SEER 18 Regs Public Use.

6. North American Association of Central Cancer Registries. NAACCR Incidence—CiNA Analytic File, 1995-2002.

7. Jemal A. Cancer Facts & Figures 2007. Atlanta, GA: American Cancer Society; 2007. In press.

8. Jemal A. Cancer statistics, 2007. CA Cancer J Clin 2007;57:43–66.

9. Pickle LW, Feuer EJ, Edwards BK. U.S. predicted cancer incidence, 1999: complete maps by county and state from spatial projection models. NCI Cancer Surveillance Monograph Series, number 5. Bethesda, MD: National Cancer Institute; 2003.

10. Frey CM, Feuer EJ, Timmel MJ. Projection of incidence rates to a larger population using ecologic variables. Stat Med 1994;13:1755–1770.

11. United States Cancer Statistics Working Group. United States Cancer Statistics: 1999 Incidence. Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2002.

12. United States Cancer Statistics Working Group. United States Cancer Statistics: 2000 Incidence. Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2003.

13. United States Cancer Statistics Working Group. United States Cancer Statistics: 2001 Incidence and Mortality. Atlanta, GA: Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute; 2004.

14. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Stat Med 2000;19: 335–351.

15. Joinpoint Regression Program [computer program]. Version 3.0. Washington, DC: National Cancer Institute, U.S. National Institutes of Health; 2005.

16. Howe HL, Wu X, Ries LA, et al. Annual report to the nation on the status of cancer, 1975–2003, featuring cancer among U.S. Hispanic/Latino populations. Cancer 2006;107:1711–1742.

17. Ghosh K, Tiwari RC. Prediction of U.S. cancer mortality counts using semiparametric Bayesian techniques. J Am Stat Assoc. In press.

18. SAS Institute. SAS/STAT GLIMMIX Procedure. June 2006. Available at: http://www.sas.com/apps/demosdownloads/sasstatglimmix_PROD__sysdep.jsp?packageID=000353. Accessed August 3, 2006.

19. Makuc DM, Haglund B, Ingram DD, et al. Health service areas for the United States. Hyattsville, MD: National Center for Health Statistics, Vital and Health Statistics; 1991.

20. Pickle LW, Mungiole M, Jones GK, White AA. Atlas of United States Mortality. Hyattsville, MD: National Center for Health Statistics; 1996.

21. National Cancer Institute, U.S. National Institutes of Health. Surveillance, Epidemiology, and End Results (SEER) Program. SEER Cause of Death Recode. Available at: http://seer.cancer.gov/codrecode/. Accessed November 7, 2006.

22. National Cancer Institute, U.S. National Institutes of Health. Surveillance, Epidemiology, and End Results (SEER) Program. Population Estimates Used in NCI's SEER*Stat Software. Available at: http://seer.cancer.gov/popdata/methods. Accessed November 7, 2006.

23. Office of Research and Planning, Bureau of Health Professions, Health Resources and Services Administration. Area Resource File, 1999. Available at: http://bhpr.hrsa.gov/healthworkforce/data/arf.htm. Accessed November 7, 2006.

24. United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Adult and Community Health. Behavioral Risk Factor Surveillance System, 2003. Available at: http://www.cdc.gov/brfss/technical_infodata/index.htm. Accessed November 7, 2006.

25. United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics. Mortality Data from the National Vital Statistics System. Available at: http://www.cdc.gov/nchs/deaths.htm. Accessed November 7, 2006.

26. Clegg L, Feuer EJ, Midthune DN, et al. Impact of reporting delay and reporting error on cancer incidence rates and trends. J Natl Cancer Inst 2002;94:1537–1545.

27. National Cancer Institute, U.S. National Institutes of Health. Cancer Query Systems: Delay-Adjusted SEER Incidence Rates. Available at: http://srab.cancer.gov/delay/canques. Accessed November 7, 2006.

28. National Cancer Institute. SEER Cancer Statistics Review, 1975-2003. Available at: http://seer.cancer.gov/csr/1975_2003/results_merged/sect_04_breast.pdf. Accessed November 7, 2006.

29. Huang L, Pickle LW, Stinchcomb DG, Feuer EJ. Spatial survival clusters of patients diagnosed with lung cancer and late stage colorectal cancer in California. Epidemiology. In press.

30. Waller LA, Gotway CA. Applied Spatial Statistics for Public Health Data. Hoboken, NJ: John Wiley & Sons, Inc.; 2004.