Updated Methodology for Projecting U.S.- and State-Level Cancer Counts for the Current Calendar Year: Part I: Spatio-temporal Modeling for Cancer Incidence

Benmei Liu[1], Li Zhu[1], Joe Zou[2], Huann-Sheng Chen[1], Kimberly D. Miller[3], Ahmedin Jemal[3], Rebecca L. Siegel[3], and Eric J. Feuer[1]

ABSTRACT

Background: The American Cancer Society (ACS) and the NCI collaborate every 5–8 years to update the methods for estimating numbers of new cancer cases and deaths in the current year in the United States and in every state and the District of Columbia. In this article, we reevaluate the statistical method for estimating unavailable historical incident cases which are needed for projecting the current year counts.

Methods: We compared the current county-level model developed in 2012 (M0) with three new models, including a state-level mixed effect model (M1) and two state-level hierarchical Bayes models with varying random effects (M2 and M3). We used 1996–2014 incidence data for 16 sex-specific cancer sites to fit the models. An average absolute relative deviation (AARD) comparing the observed with the model-specific predicted counts was calculated for each site. Models were also cross-validated for six selected sex-specific cancer sites.

Results: For the cross-validation, the AARD ranged from 2.8% to 33.0% for M0, 3.3% to 31.1% for M1, 6.6% to 30.5% for M2, and 10.4% to 393.2% for M3. M1 encountered the least technical issues in terms of model convergence and running time.

Conclusions: The state-level mixed effect model (M1) was overall superior in accuracy and computational efficiency and will be the new model for the ACS current year projection project.

Impact: In addition to predicting the unavailable state-level historical incidence counts for cancer surveillance, the updated algorithms have broad applicability for disease mapping and other activities of public health planning, advocacy, and research.

Introduction

In January of each year, *Cancer Facts & Figures* (1), produced by the American Cancer Society (ACS), releases projected numbers of new cancer cases and deaths for the current calendar year based on methods developed by ACS and the NCI, which are updated every 5–8 years. This currently requires a projection of 4 years ahead for incidence and 3 years ahead for mortality from the most recent available data based on the two methodology articles published in *Cancer* 2012 (2, 3). Historical cancer mortality data used for projection are available from all states. However, historical cancer incidence data are not available from every state for every year because some cancer registries did not begin operations until the 1990s, had not achieved the certification standards established by the North American Association of Central Cancer Registries (NAACCR), and/or do not release data because of state restrictions. Thus, before making projection of new cancer cases in the current year, the unavailable historical cancer case counts (referred to as missing counts hereafter) must be predicted using spatio-temporal statistical modeling approaches. This article focuses on this important spatio-temporal prediction step for filling in the missing counts during the observed data range, and the results from this step are referred as "predicted counts." In a subsequent article (K.D. Miller; submitted for publication) we describe the updated methodology for the temporal projection of current year incidence and mortality case counts, and the results are referred as "projected counts."

The key aim of spatio-temporal modeling is to account for both time and space correlations simultaneously. A variety of spatio-temporal models are developed in the literature and are popularly used in the epidemiology research. Cressie and Wikle (4) provided a full-length introduction to spatio-temporal modeling and outlined some of the standard techniques used in this area. More recently, Pickle and colleagues (2007; ref. 5) proposed county-level hierarchical Poisson mixed effect modeling of cancer incidence that incorporates potential predictors and spatial and temporal variation of cancer occurrence and account for delay in case reporting. The same method as the one proposed by Pickle and colleagues was used to produce state- and county-level maps of estimated cancer incidence in 1999 (6). This methodology has been used by ACS to fill in any missing counts in a state, and to smooth the observed case counts over time through the modeling process. The methodology was further updated in 2012 with additional covariates being added to the model (3).

Since the last update of the methodology in 2012 (3), the landscape of incidence data has changed dramatically with improved geographic coverage and quality of the data being collected (https://www.naaccr.org/cina-deluxe-for-researchers/). In addition, new and improved statistical methodologies in estimation and projection have evolved in the literature. For example, Mokdad and colleagues (2017; ref. 7) proposed hierarchical Bayesian mixed effect spatio-temporal model to estimate age-standardized mortality rates by U.S. county from 29 cancer sites. Our goal for this research is to evaluate several candidate

[1]Division of Cancer Control and Population Sciences, NCI, Rockville, Maryland. [2]Information Management Services, Inc, Calverton, Maryland. [3]American Cancer Society, Atlanta, Georgia.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

B. Liu and L. Zhu contributed as co-first authors, and R.L. Siegel and E.J. Feuer as co-senior authors to this article.

Corresponding Author: Benmei Liu, Division of Cancer Control and Population Sciences, NCI, NIH, Rockville, MD 20850. Phone: 240-276-6718; E-mail: liub2@mail.nih.gov

Cancer Epidemiol Biomarkers Prev 2021;30:1620–6

doi: 10.1158/1055-9965.EPI-20-1727

models and determine the one that predicts the missing and historical case counts most accurately for use in the annual cancer projections reported in *Cancer Facts & Figures*.

Materials and Methods

Data sources

The U.S. cancer incidence data from 1996 through 2014 for all cancer sites combined and 47 specific cancer sites (Supplementary Table S1) were obtained from the NAACCR as part of their Cancer in North America CiNA Deluxe data products (https://www.naaccr.org/cina-deluxe-for-researchers/). The 48 cancer sites include all cancer types that are available in the CiNA Deluxe database, coded per the Surveillance Epidemiology and End Results/World Health Organization site recode (8). Selected rare categories were combined, including vaginal and other female genital cancers, penile and other male genital cancers, and gallbladder and other biliary cancers. On the basis of the December 2016 submission, 30 states had complete incidence data across the years, 19 states had missing data in some years, and two states (Kansas and Minnesota) had no data available in any year (**Table 1**). Overall, about 10% of the state by year cells had no data with most missingness occurring in the earlier years.

For each cancer site, incidence counts and associated population size were stratified by sex and then tabulated by patient's year of diagnosis, age group (10 age groups), race (White, Black, and Other), and state (for all cancer sites combined and the 47 cancer-specific sites) or county (for all cancer sites combined and the 28 most common cancer sites) or Health Service Area (HSA, for the 19 rarest cancer sites). Data were tabulated by HSA for the rare cancer sites where county-level counts were too sparse (5). Partitioning HSAs (9) were used when the original HSAs crossed state borders (10). The cancer incidence counts were delay adjusted to account for expected delay in case reporting (11). Five data sources were used to extract a set of ecological covariates at the state, county and HSA level: the National Vital Statistics System (12), U.S. Census data (2000, 2010 and the American Community Survey; ref. 13), the Area Resource Files (14), and the Behavioral Risk Factor Surveillance System (BRFSS; ref. 15). Starting with the covariate list in the existing method (3), we updated the covariates with more recent time-dependent values during 1996 to 2014 for the spatio-temporal models which included 36 ecological variables and several individual-level covariates such as patient's age, sex, race, state/county and census division of residence, cancer site, square and cubic of age, and spline of year of diagnosis (**Table 2**).

Statistical models

Our goal was to predict missing cancer case counts during 1996–2014 through spatio-temporal modeling. When selecting novel approaches to compare with the current methodology, we were motivated by several considerations: (i) few data are missing in the years since the previous studies were published in 2012, (ii) the current county-/HSA-level spatio-temporal model takes considerable computation time and sometimes does not converge, and (iii) estimates are only published at the national and state level in the ACS's annual *Cancer Facts & Figures*. As such, we introduced three new models at the state level that are likely to be more stable and take less computational time to "compete" against the prior county-/HSA-level spatio-temporal model (3) through validation studies.

The prior spatio-temporal model (3), denoted as M0, was a county-level (HSA level for rare cancers) hierarchical Poisson mixed effect model. The input data were a tabulation of counts for county (or HSA) by year by age group by race. The model features the county/HSA-level analysis with fixed effects on a list of ecological covariates, along with random effect components including a spatial random effect, a temporal random effect, and a residual random effect, to account for any remaining "overdispersion," or greater than expected variation in the model. The major advantage of this method is the model construction at the finer geographic scale of county or HSA. The county-level model benefits from the "borrow of strengths" of the ecological predictors at the finer geographic scale (compared with the state-level analysis). This was especially important in the earlier years when many state registries had just started submitting data to NAACCR. A relatively larger proportion of registries failed the data quality certification in the earlier years, and hence there were more missing data and the spatial coverage was not as comprehensive as it is now. The disadvantage, however, is mainly the manual fine tuning of the random effect parts to achieve convergence and the computational burden.

In recent data years, cancer registration coverage is more complete and there are fewer missing state-years, allowing for the development and evaluation of a state-level method, labeled M1. In M1, we aggregated the county/HAS-level cancer counts and ecological predictors to the state level and kept the same random effect components as in M0. The input data reduced to a tabulation of counts for state by year, age group, and race. Model M1 greatly reduced the computational burden while achieving a certain level of automated model running for improved practicality. Both models M0 and M1 were implemented in SAS PROC GLIMMIX (16).

Motivated by the Bayesian model proposed in recent work (7), and to account for possible autocorrelation among states and interactions between states, time and age, we developed a state-level hierarchical Bayes spatio-temporal Poisson mixed effect model—denoted as M2. Model M2 adopted the same fixed effects on the ecological predictors as in M1, but the random effect components are more complex, consisting of a random race effect, a spatial effect for each state, two space–time interaction terms, two space–age interaction terms, a time–age interaction term, and a residual random effect. Distributions accounting for autocorrelation between neighboring states were assumed to be the form of the three spatial-related random effects. Independent mean-zero normal distributions were assumed to be the form of the five remaining random effect terms.

Given the complexity of M2, we also explored whether a simpler model could perform equally well or better. We thus introduced a simplified hierarchical Bayesian spatio-temporal Poisson mixed effect model—denoted as M3, by including only the spatial random effect for each state, a temporal random effect for each year, and a residual random effects term. Distributions accounting for autocorrelation between neighboring states were assumed to be the form of the spatial random effects and independent mean-zero normal distributions were assumed to be the form of the temporal and residual random effects. The formulation of M3 was similar to the general hierarchical Bayes spatio-temporal models studied in the literature (17, 18). Both models M2 and M3 had the same input data structure as model M1 and were implemented using fully Bayesian approach through SAS PROC MCMC (19).

Before running those models, classical model selection procedures (SAS logistic regression with backward selection using cancer incidence rates with a logarithmic transformation as the dependent variable) were applied to reduce the number of auxiliary variables. See the online Supplementary Materials and Methods for more technical details on the four models and how they were implemented.

Table 1. NAACCR CiNA incidence missing from years 1996–2014 based on the December 2016 submission.

State	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Alabama	x	x																	
Arkansas	x	x	x																
District of Columbia	x						x												
Georgia	x																		
Kansas	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x
Maryland	x	x	x																
Massachusetts	x																		
Minnesota	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x
Mississippi	x	x	x	x	x	x	x	x											
Missouri	x	x																	
Nevada	x															x	x	x	x
New Hampshire	x	x	x																
New Mexico																		x	x
North Dakota	x	x																	
Ohio	x																		
Oklahoma	x																		
South Carolina	x																		
South Dakota	x	x	x	x	x														
Tennessee	x	x	x	x	x	x	x	x											
Virginia	x	x	x	x	x	x													
West Virginia	x																		

Note: "x" means incidence data are not available (no data or not fit for use) in the CiNA Deluxe. The remaining 30 states that are not included in this table had no missing data in the CiNA deluxe.

Model evaluation

Model evaluation using original data

We ran each of the four models on the original data. To measure the accuracy of the prediction counts using each model, the prediction error was defined as the difference between the predicted incidence counts and the corresponding delay-adjusted observed incidence counts aggregated to states (even in model M0 which is a county/HSA level model) by year of diagnosis (all races combined and by race). The absolute relative deviation (ARD) was defined as the ratio of the absolute prediction error to the observed counts plus a small positive constant c for each state s and each year of diagnosis t, that is,

$$ARD_{st} = \frac{|Predicted_{st} - Observed_{st}|}{Observed_{st} + c}, \mathrm{s} = 1, \ldots, \mathrm{S}, t = 1, \ldots, T.$$

Adding an arbitrary small positive constant c ($c = 0.01$) to the observed counts was done to avoid a division by 0. To evaluate the performance of the different spatio-temporal models, we further computed the average ARD (AARD) and the median ARD (MARD) over all the states and years of diagnosis in the data for a specific cancer site. That is, $AARD = \frac{1}{ST}\sum_{s=1}^{S}\sum_{t=1}^{T} ARD_{st}$, and $MARD = Median(ARD_{st,}\ \mathrm{s} = 1, \ldots, \mathrm{S},\ t = 1, \ldots, T)$. AARD is interpreted as the average percentage deviation of the predicted values from the observed data. The MARD measure is used to account for some outliers in the average relative deviation which occurred when the denominator was very close to zero.

It would be too overwhelming to evaluate the four models using all the 47 specific cancer sites plus all cancer sites combined by sex. Thus, for this evaluation, we examined 16 randomly selected sex-specific cancer sites representing a wide range of cancer sites from common to rare cancers (listed in **Table 3**).

Cross-validation using simulated data

Cross-validation using simulated data were needed to see how well each model performed in terms of predicting missing counts. We could only do a limited cross-validation due to computation time and convergence issues. Because there were about 10% of the state by year cells with missing incidence during 1996–2014, we used 10-fold cross-validation. We split the input data into 10 random groups, with each group having approximately the same number of aggregated state by year cells. We then set all of the observed incidence from the first random group, which would later be treated as the "true" values for the cross-validation, to missing. We next applied all four spatio-temporal models to the overall data (both the first random group, now set entirely to missing, and the remaining data as is) to predict the missing incidence in the first random group, including both the original missing cells and those that were set to missing on purpose. This process was repeated for all 10 cross-validation groups. Thus, each spatio-temporal model was repeated 10 times, with the input data varying only by the artificially set to missing data in the incidence variable. As an important note, we included the actual missing data (i.e., the 10% missing state by year cells) throughout the cross-validation because the spatio-related random effects in models M2 and M3 required autocorrelation information from all the 50 states and District of Columbia, and deleting those original missing data would remove two states entirely.

The prediction results for all the missing values which had actual observed (true) values were combined from the 10 runs for each model. Then both the predicted incidence and the corresponding "true" were aggerated to the state by year level. The ARD was calculated from the predicted values and the corresponding "true" values for each model and analyzed through boxplots. AARD and MARD were also calculated from those ARDs.

Because of the intensive computational burden for this cross-validation, instead of examining all the 16 selected cancer sites, we examined six selected cancer sites varying from rare to common cancers including female all cancer combined, female lung and

Table 2. The pool of covariates from years 1996 to 2014 and data sources.

Variables	Data sources
Personal characteristics	
Year (and spline of year)	NAACCR CiNA Deluxe
Age (10 groups for common cancer: 0–4, 5–14, 15–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, 85+), age square, age cubic	NAACCR CiNA Deluxe
Race (W/B/O)	NAACCR CiNA Deluxe
Census divisions (9)	NAACCR CiNA Deluxe
Ecological variables	
State FIPS	NAACCR CiNA Deluxe
County FIPS	NAACCR CiNA Deluxe
HSA ID	NAACCR CiNA Deluxe
Mortality rate (by state, age and race and year)	NAACCR CiNA Deluxe
State-level, county-level, and HSA-level ecological variables	
% of people ages under 18	Census & American Community Survey
% of people ages 65 plus	Census & American Community Survey
% of Hispanic	Census & American Community Survey
% of American Indian Alaska Native	Census & American Community Survey
% of Black	Census & American Community Survey
% of White	Census & American Community Survey
% of persons who are foreign born	Census & American Community Survey
% of households that are linguistically isolated	Census & American Community Survey
% of households headed by female	Census & American Community Survey
% of households with more than one person per room	Census & American Community Survey
% of people living in rural areas	Census
Population density (# persons/square mile)	Census
% of people ages 25+ with <9 years of education	Census & American Community Survey
% of people ages 25+ with at least bachelor's degree	Census & American Community Survey
% of families whose incomes are below poverty	Census & American Community Survey
% of persons whose incomes are below poverty	Census & American Community Survey
% of persons ages 16+ who are unemployed	Census & American Community Survey
% of persons ages 16+ employed in white collar jobs	Census & American Community Survey
Median household (or family) income	Census & American Community Survey
Density of medical doctors (# MD/1,000 pop)	Area Resource File
Density of medical facilities (# units/1,000 pop)	Area Resource File
Land area in square miles	Census geographic file
Latitude	Census geographic file
Longitude	Census geographic file
State-level ecological variables	Sources
% of persons ages 18+ who do not have a health plan or health insurance	Area Resource File
% of females ages 18+ who ever smoked cigarettes	BRFSS
% of males ages 18+ who ever smoked cigarettes	BRFSS
% of persons ages 18+ whose BMI ≥25	BRFSS
% of persons ages 50–74 who have conducted a high-sensitivity FOBT at home in the past year or have reported having at least one colorectal endoscopy (proctoscopy, sigmoidoscopy, or colonoscopy) in his/her life	BRFSS
% of women ages 50–74 who had a mammogram in past 2 years	BRFSS
% of women ages 21–65 who had a Pap smear in past 3 years	BRFSS
% of men ages 50+ who have ever had a PSA test	BRFSS

Note: For data that came from Census only, Census 2000 estimate was used for years 1996–2005 and Census 2010 was used for 2006–2014. For data that came from Census or American Community survey (ACS): Census 2000 estimate was used for years 1996–2005. For years 2006–2014, ACS 1-year state estimate was used for state-level data; ACS 5-year county average estimate was used for county-level and HSA-level data. We labeled the ACS 5-year average statistics at the midpoint: 2006–2010 as 2008, 2007–2011 as 2009, etc. We used ACS 2006–2010 average for years 2006, 2007, and 2008 and used 2011–2015 for 2013 and 2014.

bronchus, male melanoma of skin, female brain and other nervous system, male acute lymphocytic leukemia, and female eye and orbit.

Results

Model evaluation using original data

Table 3 presents the overall summaries of the model evaluation results. The sum of the 1996–2014 incidence counts prior to filling in the missing counts varied from 12,480,762 to 21,575, the AARD ranged from 3.0% to 32.4% for M0, 2.0% to 28.0% for M1, 0.1% to 26.2% for M2, and 0.1% to 25.0% for M3 from more common cancer to rarer cancer across the 16 selected sex- and cancer-specific sites. The Pearson correlation coefficients of the AARDS of M1, M2, M3 with M0 are 0.992, 0.969, and 0.968 ($P < 0.0001$). The overall AARD and MARD indicated that for the state by year groups with observed incidence, the predicted values from the two MCMC models (M2 and M3) were

Table 3. The summary relative deviation between predicted and observed incidence counts (across all state by year groups with observed incidence).

Cancer site and sex	Sum of the 1996–2014 incidence counts prior to prediction of the missing counts	AARD (%)[a]				MARD (%)			
		M0	M1	M2	M3	M0	M1	M2	M3
All sites, female	12,480,762	3.1[b]	2.0	0.0	0.1	2.2[b]	1.5	0.0	0.0
Breast, female	3,720,450	3.0	3.1	1.0	0.9	2.2	2.2	0.4	0.4
Prostate, male	3,664,555	7.7	6.3	0.1	0.2	6.0	4.8	0.1	0.1
Lung and bronchus, female	1,666,824	4.2	3.7	2.1	1.9	3.3	2.5	1.0	0.9
Melanoma of the skin, male	584,759	10.2	9.2	3.6	5.6	8.0	6.9	1.6	3.3
Thyroid, female	436,292	8.5	8.9	4.6	5.8	5.7	6.6	2.2	3.2
Myeloma, male	186,518	9.7	9.6	8.3	9.2	6.6	6.8	5.3	6.2
Uterine cervix, female	234,374	9.0	9.6	7.7	6.4	6.2	6.2	4.5	3.2
Brain and other nervous system, female	164,996	9.3[b]	9.2	8.6	8.8	6.5[b]	6.6	5.7	6.1
Stomach, female	149,724	11.1	11.2	10.5	10.6	7.3	6.6	6.2	5.9
Mouth, male	115,419	11.7	12.0	10.8	11.1	8.0	8.0	6.4	6.7
[c]Hodgkin lymphoma, male	81,354	13.4	13.7	12.3	11.8	8.9	9.6	7.8	7.4
[c]Acute lymphocytic leukemia, Male	46,048	21.9	21.6	20.6	20.8	12.5	11.9	11.1	11.1
[c]Penis and other genital, Male	26,993	25.7	25.7	24.3	24.8	15.2	14.6	13.0	13.4
[c]Bones and joints, male	28,306	27.0	26.5	25.7	24.4	15.1	14.8	14.1	12.8
[c]Eye and orbit, female	21,575	32.3	28.0	26.2	24.9	19.6	16.5	15.0	13.5
Median	175,757	10.0	9.7	8.5	9.0	7.0	6.7	5.5	6.0

Note: M0: Previous county-/HSA-level hierarchical Poisson mixed effect model implemented in SAS GLIMMIX.
M1: State-level hierarchical Poisson mixed effect model implemented in SAS GLIMMIX.
M2: State-level complex hierarchical Bayes spatio-temporal Poisson mixed effect model implemented in SAS PROC MCMC.
M3: State-level simplified hierarchical Bayes spatio-temporal Poisson mixed effect model implemented in SAS PROC MCMC.
[a]The Pearson correlation coefficients of the AARDs of M1, M2, and M3 with M0 are 0.992, 0.969, and 0.968 ($P < 0.0001$).
[b]Spatial random effect was removed from the M0 model to make the model converge.
[c]Model M0 was run at the HSA level instead of county level.

generally closer to the corresponding observed values compared with those from the two GLIMMIX models (M0 and M1). The performance of M2 and M3 was similar, with M2 outperforming M3 for 10 of the 16 cancer sites. M1 and M0 performed similarly for most of the cancer sites, with M1 performing better for 10 of the 16 cancer sites. Bigger performance differences were detected among the models for more common cancers compared with rarer cancers. In addition, all the models performed better for more common cancers than for rarer cancers as the data became sparser in rarer cancers.

Cross-validation using simulated data

The median sum of the 1996–2014 incidence counts prior to filling in the missing counts from the 10 simulated data varied from 11,240,015 for female all cancer combined to 19,430 for female eye cancer. All four models encountered some level of technical issues such as nonconvergence or failure to run. M1 encountered the least number of issues (2 runs), while M0 and M3 encountered the most (14 runs).

Figure 1 presents the boxplots of the ARD between predicted and "true" incidence counts from each model for each cancer site. **Table 4** displays the corresponding overall AARD and MARD. M1 and M0 performed similarly, except that M1 always had lower extreme values in the ARD and had less convergence issues than M0. M1 and M0 visibly outperformed M2 and M3 for female all cancer combined and for female brain and other nervous system. M3 performed worse than the other models, with substantially higher ARDs for female all cancer types combined and female eye (**Fig. 1**).

Discussion

During the 8 years since the last update of the statistical methodologies being used by the ACS' *Cancer Facts & Figures* for projecting the current year cancer cases and deaths, statistical methods of estimation have advanced and many factors that impact cancer incidence have changed, such as growth of the population, improvements in data collection techniques and data quality, and the prevalence of risk factors. This article focused on reevaluating the spatio-temporal models to fill in the missing historical cancer incidence data in the years prior to the 4-year ahead projection, the first step for projecting current year cancer cases.

We focused on state-level modeling rather than county-level modeling for this evaluation, given that the ACS only releases national-level projections for each of the 47 cancer sites and state-level projections for selected cancer sites (1), and that county-level data are much more sparse than state-level data. The comparisons between predicted and observed cancer incidence in terms of AARD and MARD showed obvious improvements using state-level models over the original county-level model. We initially expected that the complex hierarchical Bayes model (M2) implemented through SAS PROC MCMC would perform the best because it accounts for all possible random effects; however, the cross-validation study indicated that the performance of M2 varies depending on the input data, and this type of models may not be simply replicated to other datasets because each model fit requires careful model diagnosis for convergence and goodness-of-fit. This is a major computational disadvantage since 80+ models must be run separately, one for each cancer- and sex-specific site, and an automation process must be used. Although the state-level mixed effect model (M1) implemented through SAS PROC GLIMMIX was not the best model in the initial evaluation, it performed the best in the cross-validation study and outperformed the two hierarchical Bayesian models for half of the six cancer types. On the basis of the evaluation studies and practical application concerns, model M1

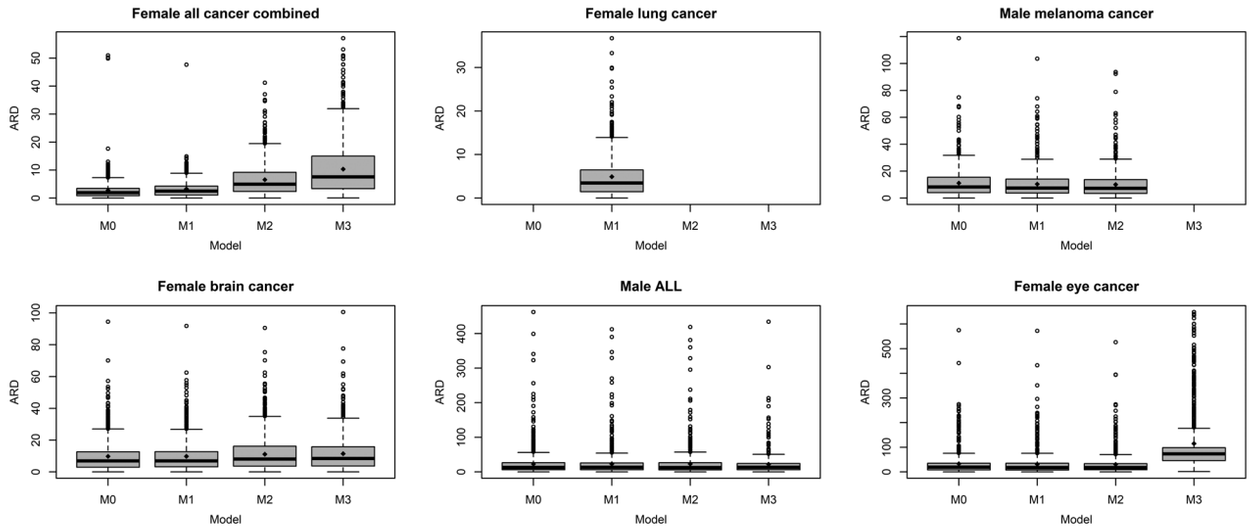

Figure 1.
Boxplots of the ARD by model from the cross-validation (M0, M2, and M3 did not converge for female lung cancer; M3 did not converge for male melanoma cancer).

was selected as the one for use starting with Cancer Facts & Figures for 2021.

Only M0 encountered nonconvergence issues in the initial evaluation which used the original data. However, all four models encountered different levels of nonconvergence issues in the cross-validation study when about an additional 10% missing data was introduced in the simulations. Although M1 had the fewest convergence issues in the cross-validation, future problems may still arise for a few cancer sites depending on the input data. One practical alternative is to alter the model slightly by removing one (or both) of the spatial or temporal random effects until the model converges, a practice that has been used by ACS in the past (1). We expect this issue to occur less often with the use of M1 than it had using the original county-level model (M0).

All models performed better for more common cancers than for rarer cancers likely reflecting the impact of sparse data. For example, for the state-level data (state by sex, race, age group, and year), among the cells with observed incidence, the percentage of cells with observed zero incidence increased from 9.4% for female all-cancer combined to 76.9% for female eye cancer site. The original county-level model and the three state-level models all assumed a Poisson distribution for the observed incidence counts in the level 1 of the models. This assumption is reasonable for the common cancer sites. However, for rare cancer sites, it may have been more suitable to assume a zero-inflated Poisson, which allows for overdispersion. We will investigate those assumptions and the practicality of doing so for future research.

One limitation of our study is that we only evaluated the models on a sampled set of cancer sites instead of all the 47 cancer sites due to computational intensity. An additional limitation is that because we took out an additional 10% observed counts as "missing" in the cross-validation step, the increased percent of missing would have a greater impact on the Bayesian hierarchical models because those models rely more heavily on the prior distributions. One potential remedy to reduce the impact was to exclude the original missing counts from the

Table 4. The summary relative deviation between predicted and "true" incidence counts (cross-validation).[a]

Cancer site and sex	Median sum of the 1996–2014 incidence counts from the 10 simulated data prior to prediction of the missing counts	AARD				AARD			
		M0	M1	M2	M3	M0	M1	M2	M3
All sites, female	11,240,015	2.8	3.3	6.6	10.4	1.9	2.5	5.0	7.6
[a]Lung and bronchus, female	1,498,806	—	4.9	—	—	—	3.5	—	—
[a]Melanoma of the skin, male	525,285	11.2	10.4	10.1	—	8.3	7.5	7.3	—
Brain, female	148,433	9.8	9.8	11.1	11.5	7.0	6.9	8.2	8.4
[b]Acute lymphocytic leukemia, male	41,388	22.9	23.0	23.1	22.1	13.2	13.5	12.9	13.7
[b]Eye and orbit, female	19,430	33.0	31.1	30.5	393.2	19.5	18.4	18.6	81.6
Median	336,859	11.2	10.1	11.1	16.8	8.3	7.2	8.2	11.1

Note: M0: Previous county-/HSA-level hierarchical Poisson mixed effect model implemented in SAS GLIMMIX.
M1: State-level hierarchical Poisson mixed effect model implemented in SAS GLIMMIX.
M2: State-level complex hierarchical Bayes spatio-temporal Poisson mixed effect model implemented in SAS PROC MCMC.
M3: State-level simplified hierarchical Bayes spatio-temporal Poisson mixed effect model implemented in SAS PROC MCMC.
[a]From the 10 runs for each cancer site each model, the following models failed to converge. Model M0: 1 run of female all cancer combined, 1 run of male melanoma, 2 runs of male eye cancer, and all 10 runs of female lung; Model M1: 2 runs of female all cancer combined. Model M2: all 10 runs of female lung; Model M3: 4 runs of male acute lymphocytic leukemia, all 10 runs of female lung, and all 10 runs of male melanoma.
[b]Model M0 was run at the HSA level instead of county level.

cross-validation, so the percent of missing counts in the simulated data was kept at 10%. However, as we explained earlier, this approach is not possible because we had to keep the original missing data because the two whole states that had no data (Kansas and Minnesota) had to be included in the spatial random effect constructions. In part II of this project (K.D. Miller; submitted for publication), we evaluated the impact of filling in the missing data by considering two sets of input: modeled data and only modeled when observed data were missing.

Similar to the models that were studied in the literature (7, 17, 18), the models studied in this article have potential broad applicability such as disease mapping and other applications. The final results from this project (Part I & II) are available in *Cancer Facts & Figures*, 2021 (https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/). These estimates will be continue to be used to fill in important gaps in cancer surveillance and other public health applications.

Authors' Disclosures

J. Zou reports personal fees from NCI during the conduct of the study, as well as personal fees from NCI outside the submitted work. R.L. Siegel is employed by the American Cancer Society, which receives grants from private and corporate foundations, including foundations associated with companies in the health sector for research outside of the submitted work. R.L. Siegel is not funded by or key personnel for any of these grants and her salary is solely funded through American Cancer Society funds. No disclosures were reported by the other authors.

Authors' Contributions

B. Liu: Conceptualization, formal analysis, validation, methodology, writing–original draft. **L. Zhu:** Conceptualization, formal analysis, validation, methodology, writing–original draft. **J. Zou:** Data curation, methodology, writing–review and editing. **H.-S. Chen:** Methodology, writing–review and editing. **K.D. Miller:** Methodology, writing–review and editing. **A. Jemal:** Methodology, writing–review and editing. **R.L. Siegel:** Supervision, methodology, writing–review and editing. **E.J. Feuer:** Conceptualization, supervision, methodology, writing–review and editing.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 5, 2021; revised March 17, 2021; accepted May 27, 2021; published first June 22, 2021.

References

1. American Cancer Society. Cancer facts & figures 2020. Atlanta: American Cancer Society; 2020.
2. Chen HS, Portier K, Ghosh K, Naishadham D, Kim H-J, Zhu L, et al. Predicting US- and state-level cancer counts for the current calendar year: Part I: evaluation of temporal projection methods for mortality. Cancer 2012;118:1091–9.
3. Zhu L, Pickle LW, Ghosh K, Naishadham D, Portier K, Chen H-S, et al. Predicting US- and state-level cancer counts for the current calendar year: Part II: evaluation of spatiotemporal projection methods for incidence. Cancer 2012; 118:1100–9.
4. Cressie N, Wikle C. Statistics for spatio-temporal data. Hoboken, NJ: Wiley; 2011.
5. Pickle LW, Hao Y, Jemal A, Zou Z, Tiwari RC, Ward E, et al. A new method of estimating United States and state-level cancer incidence counts for the current calendar year. CA Cancer J Clin 2007;57:30–42.
6. Pickle L, Feuer EJ, Edwards BK. U.S. Predicted Cancer Incidence, 1999: complete maps by county and state from spatial projection models. Bethesda, MD: NCI; 2003.
7. Mokdad AH, Dwyer-Lindgren L, Fitzmaurice C, et al. Trends and patterns of disparities in cancer mortality among US counties, 1980–2014. JAMA 2017;317: 388–406.
8. NCI. Surveillance, epidemiology, and end results program, site recode. Available from: https://seer.cancer.gov/siterecode.
9. Makuc DM, Haglund B, Ingram DD, Kleinman JC, Feldman JJ. Health service areas for the United States. Vital Health Stat 2 1991:1–102.
10. NCI. Health service areas (HSA). Available from: http://seer.cancer.gov/seerstat/variables/countyattribs/hsa.html.
11. NCI. Development of the delay model. Available from: http://surveillance.cancer.gov/delay/model.html.
12. National Center for Health Statistics. Mortality data. Available from: http://www.cdc.gov/nchs/deaths.htm.
13. US Census Bureau. Explore census data. Available from: https://data.census.gov/cedsci/.
14. Heath Resources & Services Administration. Area Health Resources Files. Available from: https://data.hrsa.gov/topics/health-workforce/ahrf.
15. Centers for Disease Control and Prevention. Behavioral risk factor survrillance system. Available from: https://www.cdc.gov/brfss/index.html.
16. SAS Institute Inc. The GLIMMIX procedure. Available from: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_glimmix_syntax.htm&docsetVersion=14.3&locale=en.
17. Lawson A. Bayesian disease mapping: hierarchical modeling in spatial epidemiology. Boca Raton, FL: Chapman & Hall/CRC Press; 2013.
18. Khana D, Rossen LM, Hedegaard H, Warner M. A Bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with R-Inla. J Data Sci 2018;16:147–82.
19. SAS Institute Inc. The MCMC procedure. Available from: https://documentation.sas.com/?docsetId=statug&docsetTarget=statug_mcmc_syntax01.htm&docsetVersion=14.3&locale=en.