HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2018 May 03.

Published in final edited form as:

Stat Med. 2013 September 30; 32(22): 3911–3925. doi:10.1002/sim.5833.

Predicting county-level cancer incidence rates and counts in the United States

Binbing Yu

Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

Abstract

Many countries, including the United States, publish predicted numbers of cancer incidence and death in current and future years for the whole country. These predictions provide important information on the cancer burden for cancer control planners, policymakers and the general public. Based on evidence from several empirical studies, the joinpoint (segmented-line linear regression) model has been adopted by the American Cancer Society to estimate the number of new cancer cases in the United States and in individual states since 2007. Recently, cancer incidence in smaller geographic regions such as counties and FIPS code regions is of increasing interest by local policymakers. The natural extension is to directly apply the joinpoint model to county-level cancer incidence data. The direct application has several drawbacks and its performance has not been evaluated. To address the concerns, we developed a spatial random-effects joinpoint model for county-level cancer incidence data. The proposed model was used to predict both cancer incidence rates and counts at the county level. The standard joinpoint model and the proposed method were compared through a validation study. The proposed method out-performed the standard joinpoint model for almost all cancer sites, especially for moderate or rare cancer sites and for counties with small population sizes. As an application, we predicted county-level prostate cancer incidence rates and counts for the year 2011 in Connecticut.

Keywords

Cancer incidence; Joinpoint model; spatial correlation; SEER; ZIP model

1. Introduction

Cancer is the second-leading cause of death among Americans. One out of every four deaths in the United States (US) is due to cancer. The American Cancer Society (ACS) estimated that in 2011, about 1,596,670 Americans would receive a new diagnosis of invasive cancer, and 571,950 Americans would die of this disease [1]. These estimates did not include in situ cancers or the more than 1 million cases of basal and squamous cell skin cancers. According to a recent report from the Centers for Disease Control and Prevention (CDC) [2], the cost of cancer care in the US has nearly doubled in the past 20 years, and the rising costs were mainly driven by the increase in cancer prevalence, not cost per patient. The National Cancer Institute (NCI) estimated that the cost of cancer in the year 2020 is projected to reach at least $158 billion (in 2010 dollars) [3]. These figures do not include other types of costs, such as lost productivity, which add to the overall financial burden of cancer. These costs are likely

to increase because of the anticipated growth and aging of the US population. Population-based cancer statistics are crucial for health planners, policymakers, and cancer information providers in order to prioritize investments in cancer control and prevention. Researchers may use cancer statistics to investigate the effect of cancer control planning, to characterize the heterogeneity of geographical areas and demographic groups, and to examine health disparities among different groups. The ACS has published predicted numbers of cancer incidence and death per year for the whole US and individual states in the annual publication *Cancer Facts and Figures* [1] since 1960 and in *Cancer Statistics* [4] since the early 1970s. In a joint effort by NCI, ACS, CDC and the North American Association of Central Cancer Registries, Inc., the *Annual Report to the Nation on the Status of Cancer, 1975–2006* [5] provides an update on trends in cancer incidence and death rates in the US. This report also includes trends in colorectal cancer incidence and death rates and highlights the use of micro-simulation modeling as a tool for interpreting past trends and projecting future ones.

Cancer incidence is one of the most important measures of cancer burden. The cancer incidence count is the number of newly-diagnosed cancer cases of a specific site occurring in the population. The cancer incidence rate is the ratio of cancer incidence count to population count, usually expressed as the number of cases per 100,000 population at risk. Both the incidence rate and the incidence count are important from the perspective of public health. Cancer incidence counts are helpful in determining cancer burden and specific needs for services for a given population. Cancer incidence rates, on the other hand, can be used to evaluate trends and effects of cancer control and prevention. The statistical methods of predicting cancer incidence rates and counts have evolved over the years. Various methods, e.g., the projection method [6], the state-space model [7], the age-period-cohort model [8] and the generalized additive model [9], have been used to predict cancer incidence and death in the US and other countries. Kim et al. [10] applied the joinpoint models (JPM) to cancer incidence rates and proposed a permutation test to determine the optimal number of joinpoints. As a special case of spline regression models, the JPM is ideal for the interpretation of cancer trends [4; 5; 11; 12; 13; 14].

So far, use of the JPM has been restricted to cancer incidence and mortality data for the whole country and individual states. For local policymakers and health researchers, it is important to seek detailed county-level cancer incidence data to assess the burden of cancer in local regions. Several state and local governments have started to estimate current and future county-level cancer incidences. See Cooper et al. [15] for medical claims data and *Illinois County Cancer Statistics Review Incidence, 1997–2001* [16] and *Texas Cancer Facts and Figures 2008* [17] for population-based cancer incidence data. Most estimations and predictions for county-level data are limited to descriptive analysis and simple linear regression, however. As an extension, the JPM can be used to the predict county-level cancer incidence. However, there are a few challenges in using county-level data. First, county-level cancer incidence data tend to have large variation and depend heavily on county population size. Second, it is important to incorporate both temporal trends and spatial correlations among multiple counties in the same geographical area. To address these issues, we developed a spatial random-effects joinpoint (SRJP) model, in which a JPM is used for temporal cancer incidence trends and random effects are utilized to incorporate correlations of spatially adjacent counties. The possibility of using a zero-inflated Poisson (ZIP) model

[18] with excess zeros for rare cancers was also explored. The performance of the proposed model and of the standard JPM was evaluated by a validation study using population-based cancer registry data.

The rest of the paper is organized as follows. In Section 2, county-level cancer incidence data are described and the potential problems of using the JPM are illustrated with several real examples. The SRJP model for county-level cancer incidence rates is presented in Section 3. A validation study is carried out to compare performance of the two methods in Section 4. As an application, the proposed method is applied to predict county-level prostate cancer incidence data in Connecticut in Section 5. The paper ends with a discussion in the final section.

2. Data and Motivation

2.1. County-level cancer incidence data from the SEER Program

In order to answer key questions about cancer burden and cancer-related health status in diverse regions and populations in the US, the Surveillance, Epidemiology, and End Results (SEER) Program (http://www.seer.cancer.gov) was launched by NCI in 1974. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28% of the US population. The SEER Program is the only comprehensive source of population-based information in the US. The mortality data reported by SEER are provided by the National Center for Health Statistics (NCHS). Population data used in calculating cancer incidence rates are obtained periodically from the US Census Bureau.

The annual SEER publication *Cancer Statistics Review* (CSR) contains incidence, mortality, prevalence, and survival statistics from 1975 through the most recent year for which data are available. There is a four year gap between the current calendar year and the most recent year with available data. For example, if the current calendar year is 2011, the most recent year with available cancer diagnosis data would be 2007. The long-term incidence trends and survival data for this report are from five states (Connecticut, Hawaii, Iowa, New Mexico, and Utah) and four metropolitan areas (Detroit, Atlanta, San Francisco-Oakland, and Seattle-Puget Sound); this set of registries is called SEER 9. In this article, we focus on four states with multiple counties, i.e., Connecticut, Iowa, New Mexico and Utah. In total, there are 168 counties with 8, 99, 32 and 29 counties from the states of Connecticut, Iowa, New Mexico and Utah, respectively. We consider the county-level incidence data from 1982 to 2007 for all cancer sites combined and 20 individual cancer sites for males and females, separately. The mean, median and maximum of annual cancer incidence rates and counts for the 168 counties under study are presented in Table 1. The three most common cancers are prostate, lung/bronchus, and colorectal for men and breast, colorectal and lung/bronchus for women. The mean annual incidence rates are 153 per 100,000 for prostate cancer and 132 per 100,000 for female breast cancer. For individual cancer sites, both incidence rates and counts have wide ranges. The medians of county-level incidence counts for several cancer sites, e.g., esophagus and Hodgkin's lymphoma, are less than 1. The distributions of both incidence rates and counts are right skewed (data not shown). A few large counties have

extremely high incidence counts. Compared to incidence counts, incidence rates are more homogeneous and are less dependent on county population sizes.

2.2. The joinpoint regression model

The JPM [10] has been used to estimate the trend of cancer incidence rate r_i with respect to year of cancer diagnosis x_i, $i = 1, \ldots, m$. The cancer incidence rate in calendar year x_i is calculated as $r_i = d_i/n_i$, where d_i is the number of new cancer cases (incidence count) and ni is the population. The K-joinpoint model for cancer incidence rates is usually specified as

$$y_i \equiv \log(r_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \delta_k (x_i - \tau_k)^+ + \varepsilon_i, \quad (1)$$

where $(x - \tau_k)^+ = \max(0, x - \tau_k)$, $\boldsymbol{\xi} = (\beta_0, \beta_1, \delta_1, \ldots, \delta_K)$ is the vector of regression coefficients, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ is the vector of K joinpoints, and $\varepsilon_1, \ldots, \varepsilon_m$ are random errors with zero means. In the analysis of cancer incidence or mortality trends, interest often lies in the joinpoint locations τ_k's and the slope changes δ_k's. For convenience of interpretation, we assume that the joinpoint occurs at integer calendar years $\{x_1, \ldots, x_m\}$. To avoid sudden changes, we set the restrictions that a joinpoint cannot occur within L years from the first or last years in the data and that there are at least L years between the two joinpoints. For instance, for a one-JPM, the location of the joinpoint should satisfy $x_{L+1} \le \tau_1 \le x_{m-L}$. In the analysis, we set $L = 2$.

Usually one assumes that the incidence count $d_i \sim \text{Poisson}(n_i\lambda_i)$, where $\lambda_i = \text{E}(r_i)$. By Taylor's expansion, $\text{Var}(y_i) \approx 1/d_i$. Thus, when the joinpoint locations τ_k's are known, the estimates of $\boldsymbol{\xi}$ can be obtained by fitting a least square regression with weights $w_i = d_i$. When the number of joinpoints K is fixed but the locations τ_k's are unknown, the estimates of $(\boldsymbol{\tau}, \boldsymbol{\xi})$ can be obtained using the grid searching method [19]. In general, the value of K is unknown and should be selected from a plausible range $[0, K_{\max}]$. Kim et al. [20] proposed a sequential permutation-test procedure for selecting K by controlling type I error. From a Bayesian perspective, Tiwari et al. [21] recommended using the Bayesian Information Criterion for model selection. The parameter estimation and model selection methods have been implemented in the NCI-developed software *Joinpoint* (http://www.cancer.gov/srab/joinpoint). Theoretically, $K_{\max}$ could take any integer value, e.g., four or even a larger number. The computational time for a K-Joinpoint model is proportional to $\binom{m}{K}$, where m is the number of data points. In practice, a JPM with $K \ge 4$ is computationally intensive and may not be necessary in certain situations. For instance, for $m = 20$, the computation time for a 4-JPM is about $\binom{20}{4}/\binom{20}{3} = 4.25$ fold of that for a 3-JPM. Typically, $K_{\max} = 3$ is enough to capture the incidence trends for most cancer types. For example, in a recent report of cancer incidence rate trends from 1992 to 2008 [5], the maximum number of joinpoints $K_{\max} \le 2$ for all top 17 cancers for men and all top 18 cancers for women, except ovarian cancer. Therefore, we set $K_{\max} = 3$ and use the permutation test for selecting the number of joinpoints when JPM model is used.

For a specific state with J counties, we let d_{ij} and n_{ij} denote, respectively, the cancer incidence count and population count for the jth county in calendar year x_i. The corresponding county-level cancer incidence rate is calculated as $r_{ij} = d_{ij}/n_{ij}$ and the JPM assumes that

$$y_{ij} \equiv \log(r_{ij}) = \beta_{0j} + \beta_{1j}x_i + \sum_{k=1}^{K} \delta_{kj}(x_i - \tau_{kj})^+ + \varepsilon_{ij}, \quad (2)$$

where $\boldsymbol{\varepsilon}_j = (\varepsilon_{1j}, \ldots, \varepsilon_{mj})$ is the vector of random errors for the jth county. A separate set of parameters $\boldsymbol{\xi}_j = (\beta_{0j}, \beta_{1j}, \delta_{1j}, \ldots, \delta_{Kj})$ and $\boldsymbol{\tau}_j = (\tau_{1j}, \ldots, \tau_{Kj})$, $j = 1, \ldots, J$ may be used for each county. The random errors $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_J$ are independent across different counties, but $\varepsilon_{1j}, \ldots, \varepsilon_{mj}$ for the same county could be serially correlated. The JPM, as implemented in the current *Joinpoint* software, is not a true log-linear model for Poisson data. For example, the default option in *Joinpoint* software is to fit a weighted least square regression to the logarithm of incidence rate. In equation (1), a zero incidence rate is replaced by $r_i = 0.5/n_i$; if the random errors have constant variance, the whole group with at least one zero incidence will be removed from the analysis. Incidence rates for national or state-level data are usually above 0 because of the large population size. In addition to possible bias incurred by replacing zero counts with 0.5, the standard JPM does not consider spatial correlations among counties in the same geographical area. Furthermore, uncertainty of the joinpoint estimates cannot be incorporated in the prediction of incidence rates, thus under-estimating prediction errors.

2.3. Analysis and concerns

To evaluate the performance of the standard JPM, we present the analysis of county-level cancer incidence for a few selected cancers. County-level cancer incidence data from 1982 to 2007 are extracted from the SEER Program database using SEER*Stat software. Incidence data from 1982 to 2003 are used for model building and data from 2004–2007 are used for validation. For the purpose of illustration, we analyze breast cancer incidence rates and counts for the eight counties in Connecticut.

Based on equation (2), the predicted cancer incidence rates is calculated as $\hat{r}_{ij} = \exp(\hat{y}_{ij})$, where $\hat{y}_{ij}$ is the estimate of the log-incidence rate. Figure 1 shows the plot of observed and fitted cancer incidence rates. Solid dots and circles represent observed incidence rates for 1983–2003 and 2004–2007, respectively. Solid and dashed lines show the predicted trends from the proposed SRJP model and the standard JPM, respectively. Locations of the joinpoints are shown in the legends. Note that as the JPM was fitted to the logarithm of incidence rates, the predicted incidence rates were calculated by exponentiating back to the original scale. Therefore, the fitted trends have a mild curvature. We see that county-level incidence rates are homogeneous, between 80 and 180 cases per 100,000. The corresponding incidence counts for each county are shown in Figure 2. Compared to incidence rates, the county-level incidence counts are rather heterogeneous, depending heavily on the county population. For the three counties with populations greater than 400,000, i.e., Fairfield, Hartford and New Haven, the cancer incidence counts are above 500 each year, while for the

five smaller counties, the incidence counts are below 220. Both incidence rates and counts show large variations across diagnosis years, presenting a challenge for prediction.

Since the JPM is fitted separately by county, each county has its own trend. Except for the two largest counties, Fairfield and Hartford, 0-JPM is the optimal model, showing a linear increasing trend for both incidence rates and counts. The JPM tends to over-predict cancer incidence rates and counts for the years 2004–2007 for the six smaller counties at the bottom of Figures 1 and 2. For Fairfield and Hartford counties, the JPM shows an increasing trend before the year 2000 and a decline thereafter. Predictions of incidence rates from the SRJP tend to be lower than those from the JPM. The SRJP identifies two common joinpoints at 1986 and 2000. After the last joinpoint at 2000, the last segment attempts to fit the last 4 data points fairly closely and the downward trend is extrapolated to the future. This is reasonable since future predictions should be based on the most recent trends, while past trends may not contain much information about future changes. In Figure 1, predictions from the SRJP are closer to actual observed rates in 2004–2007, except for Fairfield and Middlesex counties. In Figure 2, predictions of incidence counts from the SRJP method are more accurate than from the JPM except for Middlesex county.

Therefore, the predictions from the JPM may be problematic for rare cancers or small counties where annual incidence counts could be very small or even zero. In this situation, a separate JPM for each county could break down and yield out-of-range predictions. For example, Figure 3 shows fitted trends from both the JPM and SRJP for selected cancer sites for two counties in New Mexico. For the three types of cancer in Figure 3, incidence rates are 0 for most of the years. Due to non-zero incidence in late years, the standard JPM identifies a joinpoint around the years with non-zero incidence. The resulting JPM drastically over-predicts cancer incidences in future years. This instability stems from large random fluctuation due to cancer rarity or a small county population. By incorporating spatial correlations, the SRJP should be less influenced by non-zero incidence and the corresponding predictions are much closer to observed incidence rates than are those from the JPM.

3. Statistical methods

In this section, we first present the SRJP for county-level incidence rates. Then we describe the method of predicting incidence counts by combining predicted incidence rates from the SRJP with predicted population from the US Census Bureau.

3.1. The spatial random-effects joinpoint model

We assume that county-level incidence counts $d_{ij} \sim \text{Poisson}(n_{ij}\lambda_{ij})$, and the SRJP for county-level cancer incidence rates is specified as,

$$\log(\lambda_{ij}) = \beta_{0j} + \beta_{1j}x_i + \sum_{k=1}^{K} \delta_{kj}(x_i - \tau_k)^+. \quad (3)$$

This model assumes the same joinpoints ($\tau_1, \ldots, \tau_K$) across different counties and the regression coefficients are

$$\begin{aligned}\beta_{0j} &= \mu_1 + b_{1j},\\ \beta_{1j} &= \mu_2 + b_{2j},\\ \delta_{1j} &= \mu_3 + b_{3j},\\ &\vdots\\ \delta_{Kj} &= \mu_{K+2} + b_{K+2,j},\end{aligned}$$

where $b_k = (b_{k1},\ldots,b_{kJ})$, $k = 1, \ldots, K+2$, have zero means and follow independent conditional autoregressive (CAR) distributions. The most popular CAR implementation is the intrinsic CAR model [22] with multivariate normal distribution

$$b_k \sim N(0, [\psi_k(R - W)]^{-1}),$$

where ψ_k is a scalar parameter representing the precision (inverse of variance), R is a diagonal matrix for the number of neighbors for each region, and W is the adjacency matrix of the study regions. The intrinsic CAR model is implemented in GeoBUGS, a module of BUGS [23]. The syntax for CAR distributions is

$$\text{b . k[1 : J]} \sim \text{car . normal(adj[], weight[], num[], precision . k)},$$

where `adj[]` is a vector that represents the adjacency matrix W, `weights[]` is set to 1 to be consistent with the intrinsic CAR model, `num[]` is the vector of R and `precision.k` is the precision parameter ψ_k. The intrinsic CAR model is an improper distribution and contains no parameter to control the strength of spatial dependence [24]. In situations in which spatial dependence and heterogeneity are of interest, a proper CAR distribution [25] can be used. Here, our focus is on prediction of cancer incidence rates rather than spatial dependence or pattern at county level. The intrinsic CAR model is used because it is easy to implement and has nice convergence properties.

3.2. Bayesian estimation

Let $\boldsymbol{\theta} = (\mu_1, \ldots; \mu_{K+2}, \psi_1, \ldots, \psi_{K+2}, \tau_1, \ldots, \tau_K)$ be the vector of all parameters. The joint likelihood function for county-level cancer incidence data $Y = \{x_1, \ldots, x_I; d_{i1}, \ldots, d_{iJ}; n_{i1}, \ldots, n_{iJ}; j = 1, \ldots, J\}$ can be written as

$$L(\boldsymbol{\theta} \mid Y) \propto \left(\prod_{j=1}^{J}\prod_{i=1}^{I}\frac{1}{d_{ij}!}(n_{ij}\lambda_{ij})^{d_{ij}}\exp(-n_{ij}\lambda_{ij})\right)\left(\prod_{k=1}^{K+2} N(\mathbf{0}, [\psi_k(R - W)]^{-1})\right). \quad (4)$$

where λ_{ij} is specified in equation (3).

Let $\pi(\boldsymbol{\theta})$ be the joint prior distribution of $\boldsymbol{\theta}$. The posterior distribution of $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta}|\mathrm{Y}) \propto L(\boldsymbol{\theta}|Y)\pi(\boldsymbol{\theta})$. We assume that the prior distributions for the parameters are mutually independent:

$$\pi(\boldsymbol{\theta}) = \left(\prod_{k=1}^{K+2} \pi(\mu_k)\pi(\psi_k)\right)\pi(\tau_1, \ldots, \tau_K).$$

In particular, we use conjugate priors $\mu_k \sim N(\mu_{0k}, \sigma_{0k}^2)$, $k = 1, \ldots, K+2$. There are several options for the priors of precision parameters ψ_k. The most common choice is to use a Gamma distribution ψ_k Gamma(c_k, d_k) with shape parameter c_k and scale parameter d_k. The hyperparameters, μ_{0k}, σ_{0k}^2, ck, dk, of all the above prior distributions are assumed to be known. The hyperparameters are chosen so that the priors are weakly informative. Alternatively, Gelman [26] recommended using a uniform prior for the standard deviation $\sigma_k = 1/\sqrt{\psi_k}$

A discrete prior has been used for the joinpoints [21], where the joint prior $\pi(\tau_1, \ldots, \tau_K)$ is the product of discrete uniform probability functions

$$\pi(\tau_1) = \frac{1}{m - L(K+1)}, \quad \tau_1 \in \left\{x_{L+1}, \ldots, x_{m-LK}\right\},$$
$$\pi(\tau_k | \tau_{k-1} = x_l) = \frac{1}{m - l - L(K-k)}, \quad \tau_k \in \left\{x_{l+L+1}, \ldots, x_{m-L(K-k+1)}\right\}, k \geq 2.$$

For example, the discrete uniform probability function for τ_1 is $\pi(\tau_1) = \frac{1}{m - L(K+1)}$ with support $\{x_{L+1}, \ldots, x_{m-LK}\}$.

Because the joint probability function does not have a closed form and the likelihood function is complex, one can use the Bayesian Markov chain Monte Carlo (MCMC) method to obtain the parameter estimates. The proposed method can be implemented in the freely available software WinBUGS or OpenBUGS [23]. After a sufficient number of burn-in iterations, we use the remaining samples to estimate any function of the parameters of interest. In order to see how stable the final estimates are, multiple MCMC runs are conducted with different initial values. The convergence of the MCMC samples of the parameters, after excluding the initial burn-in samples, are monitored using the R package CODA. For example, Gelman and Rubin [26] used a 'potential scale reduction factor' (PSRF) for each parameter in $\boldsymbol{\theta}$, together with upper and lower confidence limits. Approximate convergence is achieved when the upper limits are close to 1.

The deviation information criterion (DIC) [27] can be used to select the optimal number of joinpoints. The SRJP model with the smallest DIC is selected as the optimal model. Let $\boldsymbol{\theta}_K$ be the parameter set for the K-joinpoint SRJP. The DIC for the K-joinpoint SRJP is defined as

$$\mathrm{DIC}_K = \overline{D(\boldsymbol{\theta}_K)} + p_D.$$

where $D(\boldsymbol{\theta}_K) = -2\log p(Y|\boldsymbol{\theta}_K)$ is the deviance and $\overline{D(\boldsymbol{\theta}_K)}$ is the average posterior deviance, $p_D = \overline{D(\boldsymbol{\theta}_K)} - D(\overline{\boldsymbol{\theta}_K})$ is the "effective dimension", and $\overline{\boldsymbol{\theta}_K}$ is an estimate of $\boldsymbol{\theta}_K$. The DIC values from the SRJP models can be obtained directly from the output of OpenBUGS.

3.3. Sensitivity to priors

It has been criticized that parameter estimates are sensitive to the choice of Gamma distribution for the precision parameters ψ_k's in disease mapping [28]. Therefore, it is necessary to examine the sensitivity of the posterior estimates to the prior distributions. We tried different initial values and used various priors for the precision or standard deviation parameters. We used uniform priors Uniform(0.1; 100) for the standard deviation parameters σ_k and Gamma(0.1, 0.1) or Gamma(0.001, 0.001) priors for the precision parameters ψ_k. We set the hyperparameters $\sigma_{0k}^2 = 10$ or 1000. The posterior estimates of regression coefficients $\mu_1, \ldots, \mu_{K+2}$ were very close regardless of the values of σ_{0k}^2. Furthermore, the posterior estimates of $\mu_1, \ldots, \mu_{K+2}$ were similar when priors Gamma(0.1, 0,1) and Gamma(0.001, 0.001) were used for ψ_k. The uniform prior for the standard deviation σ_k led to a slightly different inference in comparison to the Gamma priors for ψ_k. In particular, the posterior mean and median of the standard deviations σ_k's were higher and the shrinkage of the μ_k's were smaller than in the SRJP with Gamma priors for ψ_k. This was consistent with the results by Gelman [26] that the uniform prior distribution on σ_k is closer to "noninformative" for the hierarchical modeling. In addition, we found that the SRJP based on the uniform prior for σ_k had slightly better fit than that based on the Gamma prior for ψ_k. Therefore, we chose the uniform prior for the σ_k's.

3.4. Predicting cancer incidence counts

Pickle et al. [12] proposed to use JPM to predict cancer incidence count at the national and state levels. Likewise, the JPM can be applied directly to county-level incidence counts d_{ij} as

$$\log(d_{ij}) = \beta_{0j}^* + \beta_{1j}^* x_i + \sum_{k=1}^{K} \delta_{kj}^* (x_i - \tau_{kj}^*)^+ + \epsilon_{ij}^*.$$

Alternatively, the cancer incidence counts can be predicted as $\widehat{d_{ij}} = n_{ij}\widehat{r_{ij}}$, where $\widehat{r_{ij}}$ is the predicted county-level cancer incidence rate based on the SRJP and n_{ij} is the county population at calendar year x_i. The population estimates are released on a flow basis throughout each year by the US Census Bureau. Each new series of data (called vintages) incorporates the latest administrative record data, geographic boundaries, and methodology. The most recent vintage year was 2011 when the analysis was done. The county population estimates that are used to calculate cancer incidence rates can be downloaded from http://www.census.gov/popest/index.html.

4. Validation Study

Using empirical data, we compared the predictive performance of the proposed SRJP model and the standard JPM. We fit the SRJP using OpenBUGS 3.1.2 and fit the standard JPM

using the software Joinpoint. The incidence data for the years 1982–2003 were used to fit the model and then the fitted model was extrapolated to the years 2004–2007. The prediction errors were calculated based on the difference between estimated and observed incidence rates/counts for the years 2004–2007. We chose data for the last four years for validation because there was a four-year lag between current calendar year and the current year with available cancer data. In total, there were 148 possible combinations of cancer site, sex and state. For each analysis, we ran two parallel chains and the posterior estimates were calculated using an additional 50,000 MCMC samples with thinning a factor of 10 after discarding the first 40,000 initial simulations. Based on the criterion of PSRF [29], the percentages of PSRF ≥ 1.1 among all sex-cancer combinations were 4%, 2% and 1% for μ_k's, σ_k's and b_{kj}'s. As the joinpoints τ's only took integer values, the PSRF values for τ's were slightly higher with 8% above 1.5. The medians of the PSRF were less than 1.02 for all the parameters. Therefore, the MCMC simulation reached convergence reasonably well. The frequencies of selecting 0–3 number of joinpoints were 122, 12, 8 and 6, respectively. This was consistent with the recent *Annual Report on the Status of Cancer* [5], which concluded that most trends of cancer incidence rates had 0 or 1 joinpoints.

4.1. Comparison of prediction accuracy

We examined the accuracy of the predictions for the years $t = 2004, \ldots, 2007$ in the 168 counties for the four states. The summary measures of prediction accuracy included the prediction error (PE), absolute prediction error (APE) and the relative prediction error (RPE), which were calculated as

$$\mathrm{PE} = \frac{1}{N_c N_t}\sum_j \sum_t (P_{jt} - A_{jt}),\quad \mathrm{APE} = \frac{1}{N_c N_t}\sum_j \sum_t |P_{jt} - A_{jt}| \text{ and } \mathrm{RPE} = \frac{1}{N_c N_t}\sum_j \sum_t \frac{|P_{jt} - A_{jt}|}{\widetilde{A}_j} \times 100\%,$$

where P_{jt} and A_{jt} were predicted and observed cancer incidence rates or counts in the year t for the jth county, $N_t = 4$ is the total number of years for prediction, $N_c = 168$ is the total number of counties. For rare cancers, the actual incidence rates may be 0, therefore the denominator in the RPE measure $\widetilde{A}_j$ was the average of the incidence rates from 1982 to 2007 for the jth county. The PE is a measure of bias indicating whether there is over- or under-predictions, which are equally undesirable. The scale of the APE may depend on the rarity of cancer at a particular site. For a rare cancer such as Hodgkin's lymphoma, the APE should be relatively small. The APE is a more meaningful measure for comparison across different cancer sites because it does not depend on rarity of cancer.

The predictions for future years from the JPM were less accurate than those from the SRJP. As illustrated in Figure 3, the predictions from the JPM sometimes drastically over-predicted cancer incidence rates. Based on Table 1, the maximum annual incidence rate was 1535 per 100,000. Therefore, predicted incidence rates above 1600 per 100,000 were considered as outliers. The prediction incidence rates from the JPM were right skewed with 119 outliers, which consisted of 0.48% of the total number of predictions. Among the 119 outlying predictions, 106 outliers only have one cancer cases. This shows that the instability of prediction is mainly because of small cancer counts in small population. After removing

these outliers, summary predictive statistics from the JPM and the SRJP are shown in Table 2. The mean PE of incidence rates from the JPM and the SRJP were 14.49 and −0.21, respectively. For the prediction of incidence counts, the difference between the two methods was smaller.

The median of prediction errors for county-level incidence rates by cancer site are summarized in Table 3. The numbers in bold indicate where the JPM had better prediction than the SRJP. Except for five cancer sites, PEs from the SRJP were smaller than those from the JPM. The JPM tended to over-predict cancer incidence rates for all cancer sites. For example, the PE from the SRJP was only −0.3 for prostate cancer, while the PE from the JPM was 27.3. The relative improvement of the SRJP was also remarkable for moderately common and rare cancers. For example, the incidence rate for pancreatic cancer was only 12.1 per 100,000 for both men and women per year. The PEs from the JPM were 5.1 and 5.5 for men and women, respectively. The corresponding PEs from the SRJP were only 0.5 and 1.4. Except for laryngeal cancer, the APE's and RPE's from the SRJP were smaller than those from the JPM. Predictions for common cancers, e.g., colorectal cancer, had larger APEs than those for relatively rare cancers, e.g., pancreatic cancer. RPE's for common cancers were smaller than those for rare cancers. This was reasonable as the incidence rates for rare cancers were very low, but with a lot of variation. Overall, predictions from the SRJP were more accurate than those from the JPM.

The prediction errors for county-level cancer incidence counts are shown Table 4. We see similar patterns that the predictions from the SRJP were uniformly better than those from the JPM. For common cancers, e.g., prostate cancer, median APE's for the JPM and SRJP were 5.5 and 3.3 and median RPE's from the JPM and SRJP were 15% and 9%, respectively. The relative improvement of the SRJP was also significant for relatively rare cancers. For example, the median APE's for the JPM and SRJP for pancreatic cancer were 1.0 versus 0.7 for men and 1.1 versus 0.8 for women, respectively. Based on the validation study, the SRJP provided much more accurate and robust predictions of incidence rates and counts than did the JPM.

4.2. Zero-inflated Poisson model

A popular alternative for cancer incidence counts with excess zeros is the Zero-inflated Poisson (ZIP) model, which is denoted as $D \sim \mathrm{ZIP}(p, \lambda)$. The probability function of a ZIP variable is given by

$$P(D = d \mid p, \lambda) = \begin{cases} p + (1-p)\exp(-\lambda) & \text{when } d = 0 \\ (1-p)\lambda^d \exp(-\lambda)/d! & \text{when } d > 0 \end{cases}. \quad (5)$$

This model assumes that cancer incidence count comes from two types of distributions. The first type gives a Poisson distributed count which might contain zeros and the second type always gives a zero count. The parameter p is the zero-inflation probability, and $\exp(-\lambda)$ is the probability of zero counts predicted by the Poisson distribution.

We assume that the county-level incidence counts $d_{ij} \sim \text{ZIP}(p, n_{ij}\lambda_{ij})$, where p is the cancer-specific zero-inflation probability, n_{ij} is the county population and λ_{ij} is specified by the SRJP (3). The resulting model is called a SRJP-ZIP model. The predicted incidence rates at year x_i for county j can be calculated as $(1-p)n_{ij}\hat{\lambda}_{ij}$. Here we used cancer-specific zero inflation probability. The probability of having zero incidence counts is $p + (1-p)\exp(-n_{ij}\lambda_{ij})$. Theoretically, one can use cancer- and county-specific zero inflation probabilities, but this may lead to too many parameters. The zero-inflation probability, i.e., the probability of having no new cancer cases, mainly depends on the rarity of cancer, not geographic locations. Therefore, it is reasonable to assume a common zero-inflation probability for all counties.

Similar to the SRJP without zero inflation, the optimal number of joinpoints was selected based on the DIC, which was provided by OpenBUGS. The PEs from the SRJP-ZIP model were found to be close to those from the SRJP model. Both the SRJP and SRJP-ZIP models had smaller prediction errors than did the JPM. For details, see supplementary Tables A and B. We did not see an improvement in prediction accuracy by using the SRJP-ZIP model for rare cancer sites. In addition, the SRJP model without zero inflation had better predictive accuracy for cancer incidence rates than did the SRJP-ZIP model for common cancer sites, for example, prostate cancer and female breast cancer. One reason for the lack of improvement of the SRJP-ZIP model was that county-level cancer incidence, especially incidence counts, were highly heterogeneous. In the SRJP, the heterogeneity could be captured by the county-level random-effects b_{kj}, $k = 1, \ldots, K+2$. As a result, the probability of having zero incidence counts was county-specific where $P(d_{ij} = 0) = \exp(-n_{ij}\lambda_{ij})$. Therefore, a small mean incidence rate λ_{ij} still provided a good fit for rare cancer sites or small counties with zero incidence counts. In addition, a zero-joinpoint SRJP tended to be selected as the optimal model. A zero-joinpoint cannot effectively capture the fluctuating variation of cancer incidence and may not provide good predictions for future years. Based on these findings, we concluded that the SRJP model was more appropriate than the usual JPM for predicting county-level cancer incidence rates and counts.

5. Application

One important use of the SRJP is to describe trends in county-level cancer incidence rates. We used the county-level prostate cancer incidence rates from Connecticut as an example. Because there are only eight counties in Connecticut, we were able to fit up to 4-joinpoint SRJP. The MCMC simulation converged quickly as the upper confidence intervals for the potential scale reduction factor R were all below 1.12. The traceplots also showed good mixing of the parameters. The DIC values for the 0–4 joinpoint SRJP were 4043, 2443, 2420, 2137 and 2255, respectively. Therefore, the 3-joinpoint model was selected as the final model. The corresponding joinpoints were the years 1988, 1992 and 1995. The first joinpoint at year 1988 reflected the use of the prostate specific antigen (PSA) screening test, which was associated with a substantial increase in the reported incidence of prostate cancer. The second joinpoint at year 1992 showed a leveling or decreasing trend in prostate cancer incidence, which may indicate the full dissemination of PSA screening. Then the incidence trends of prostate cancer incidence rates almost returned to the temporal trend after 1995.

The results showed that prostate cancer trends for Connecticut generally followed national trends [30]. Information on the magnitude of the increase, duration of time before incidence began to decline, and the eventual level at which incidence stabilized were useful to assess the impact of possible over-diagnosis and to examine the effect of lead-time bias.

The predicted prostate cancer incidence rates and counts for the eight counties in Connecticut in 2011 are shown in Table 5. The prostate cancer incidence rates are very homogeneous, with a minimum rate of 101 per 100,000 for Windham county and a maximum rate of 222 per 100,000 for Middlesex county. By combining the Census Bureau predictions of county-level population and the estimated incidence rates from SRJP, we can also predict the county-level cancer incidence counts in the year 2011. We see that two counties, Hartford and Fairfield, have the highest prostate cancer cases (above 700) in the year 2011, largely due the large population sizes in those counties. The smaller counties, Windham and Tolland, have new prostate cancer cases below 121.

6. Discussion

In this article, we developed a SRJP model to predict county-level cancer incidence rates. By combining the prediction of county-level population from the Census Bureau, the proposed method may be used to predict county-level cancer incidence counts, as well. Based on a validation study, the new method gave more accurate predictions than the current standard JPM for virtually all cancer sites and it was robust to influential observations. There are two potential issues with the proposed model, however. First, the current SRJP restricts that all counties had the same joinpoints, but allows different slopes across different counties. This is part due to the identifiability issue of the parameters. In particular, if the slopes from two consecutive segments are the same, the evidence of joinpoint is be weak. If we allow both joinpoints and slopes to vary county by county, this may cause problems in estimation. Second, the standard JPM was fitted using a frequentist method while the SRJP was fitted using a Bayesian framework. We use the standard (frequentist) JPM as the benchmark because it is the current method used by NCI for trends analysis. In this analysis, we used the Bayesian SRJP in part due to computational convenience. As a bridge, we compared the Bayesian JPM and the standard JPM for cancer incidence data at the national and state levels and found that the estimates are very similar when the priors for the variance parameters were weakly informative. This is consistent with the finding that the estimates from the Bayesian and frequentist JPM were very close [21, Table 1].

In this paper, we analyzed the crude incidence rate, which may be influenced by the underlying age distribution of the study regions. Alternatively, the age-adjusted cancer incidence rate can be used to evaluate the effect of intervention or to compare disparity among different demographic groups. The age-adjusted rate is a weighted average of age-specific (crude) rates, where the weights are the proportions of persons in the corresponding age groups of a standard population. Thus, the potential confounding effect of age is reduced by using the same standard population. The proposed SRJP model could also be used for comparative analysis using age-adjusted cancer incidence rates. The SEER Program also collects information on county-level attributes, for example, race, age, housing, education, poverty level and employment at county-level. The proposed SRJP model may incorporate

county-level characteristics as covariates to further improve predictive accuracy. In addition, the SRJP model could be used to estimate the trends of county-level cancer mortality, which is another important measure of cancer burden.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research was supported in part by the Intramural Research Program of the National Institute on Aging. The author would like to thank two anonymous referees for their insightful comments and Ms. Caroline Phillips for her conscientious editorial help and suggestions.

References

1. American Cancer Society. Cancer Facts and Figures 2011. American Cancer Society; Atlanta, Georgia, U.S.A: 2011.
2. Tangka FK, Trogdon JG, Richardson LC, Howard D, Sabatino SA, Finkelstein EA. Cancer treatment cost in the United States: has the burden shifted over time? Cancer. 2010; 116:3477–3484. [PubMed: 20564103]
3. Mariotto AB, Robin Yabroff K, Shao Y, Feuer EJ, Brown ML. Projections of the cost of cancer care in the United States: 2010–2020. Journal of the National Cancer Institute. 2011; 103(2):117–128. [PubMed: 21228314]
4. Jemal A, Siegel R, Ward E, Hao Y, Xu J, Thun MJ. Cancer Statistics, 2009. CA-A Cancer Journal for Clinicians. 2009; 59(4):225–249. [PubMed: 19474385]
5. Eheman C, Henley SJ, Ballard-Barbash R, Jacobs EJ, Schymura MJ, Noone AM, Pan L, Anderson RN, Fulton JE, Kohler BA, et al. Annual report to the nation on the status of cancer, 1975–2008, featuring cancers associated with excess weight and lack of sufficient physical activity. Cancer. 2012; 118(9):2338–2366. [PubMed: 22460733]
6. Wingo P, Landis S, Parker S. 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. 1998; 25:43–51.
7. Ghosh K, Tiwari RC. Prediction of United States cancer mortality counts using semiparametric Bayesian techniques. Journal of the American Statistical Association. 2007; 102(477):7–15.
8. Bashir S, Estève J. Projecting cancer incidence and mortality using bayesian age-period-cohort models. Journal of Epidemiology and Biostatistics. 2001; 6(3):287–296. [PubMed: 11437093]
9. Clements MS, Armstrong BK, Moolgavkar SH. Lung cancer rate predictions using generalized additive models. Biostatistics. 2005; 6(4):576–589. [PubMed: 15860544]
10. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. Statistics in Medicine. 2000; 19(3):335–351. [PubMed: 10649300]
11. Tiwari RC, Ghosh K, Jemal A, Hachey M, Ward E, Thun M, Feuer E. A new method of predicting United States and state-level cancer mortality counts for the current calendar year. CA-A Cancer Journal for Clinicians. 2004; 54(1):30–40. [PubMed: 14974762]
12. Pickle LW, Hao Y, Jemal A, Zou Z, Tiwari RC, Ward E, Hachey M, Howe HL, Feuer EJ. A new method of estimating united states and state-level cancer incidence counts for the current calendar year. CA-A Cancer Journal for Clinicians. 2007; 57(1):30–42. [PubMed: 17237034]
13. Ghosh P, Huang L, Yu B, Tiwari RC. Semiparametric Bayesian approaches to joinpoint regression for population-based cancer survival data. Computational Statistics and Data Analysis. 2009; 53:4073–4082. [PubMed: 22210971]
14. Yu B, Huang L, Tiwari RC, Feuer EJ, Johnson KA. Modelling population-based cancer survival trends by using join point models for grouped survival data. Journal of the Royal Statistical Society, Series A (Statistics in Society). 2009; 172(2):405–425.

15. Cooper GS, Yuan Z, Jethva RN, Rimm AA. Determination of county-level prostate carcinoma incidence and detection rates with medicare claims data. Cancer. 2001; 92(1):102–109. [PubMed: 11443615]
16. Qiao, B., Shen, T., Sondgrass, JL. Technical Report Epidemiologic Report Series 05:1. Illinois Department of Public Health; Jan, 2005 Illinois county cancer statistics review incidence, 1997–2001.
17. American Cancer Society, High Plains Division, Inc.. Texas facts and figures 2008. American Cancer Society, High Plains Division; 2008. Technical Report
18. Lambert D. Zero-inflated Poisson regression models with an application to defects in manufacturing. Technometrics. 1992; 34:114.
19. Lerman P. Fitting segmented regression models by grid search. Applied Statistics. 1980; 29(1):77–84.
20. Kim HJ, Yu B, Feuer E. Selecting the number of change-points in segmented line regression. Statistica Sinica. 2009; 19(2):597–609. [PubMed: 19738935]
21. Tiwari R, Cronin K, Davis W, Feuer E, Yu B, Chib S. Bayesian model selection for join point regression with application to age-adjusted cancer rates. Journal of the Royal Statistical Society, Series C (Applied Statistics). 2005; 54(5):919–939.
22. Besag J, York J, Mollie A. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics. 1991; 43(1):1–20.
23. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions (with discussion). Statistics in Medicine. 2009; 28:3049–3082. [PubMed: 19630097]
24. Cressie, NAC. Statistics for Spatial Data. 2nd. Wiley & Son; New York, NY: 1993.
25. Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: A new mixed model for spatial dependence. Statistical Models in Epidemiology, the Environment and Clinical Trials. 1999:135–178.
26. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis. 2006; 1(3):1–19.
27. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B. 2002; 64(4):583–639.
28. Eberly LE, Carlin BP. Identifiability and convergence issues for markov chain monte carlo fitting of spatial models. Statistics in Medicine. 2000; 19(17–18):2279–2294. [PubMed: 10960853]
29. Gelman A, Rubin D. Inference from alternative simulation using multiple sequences. Statistical Science. 1992; 7:457–472.
30. Feuer EJ, Merrill RM, Hankey BF. Cancer surveillance series: Interpreting trends in prostate cancer part II. Cause of death misclassification and the recent rise and fall in prostate cancer mortality. 1999; 91(12):1025–1032.

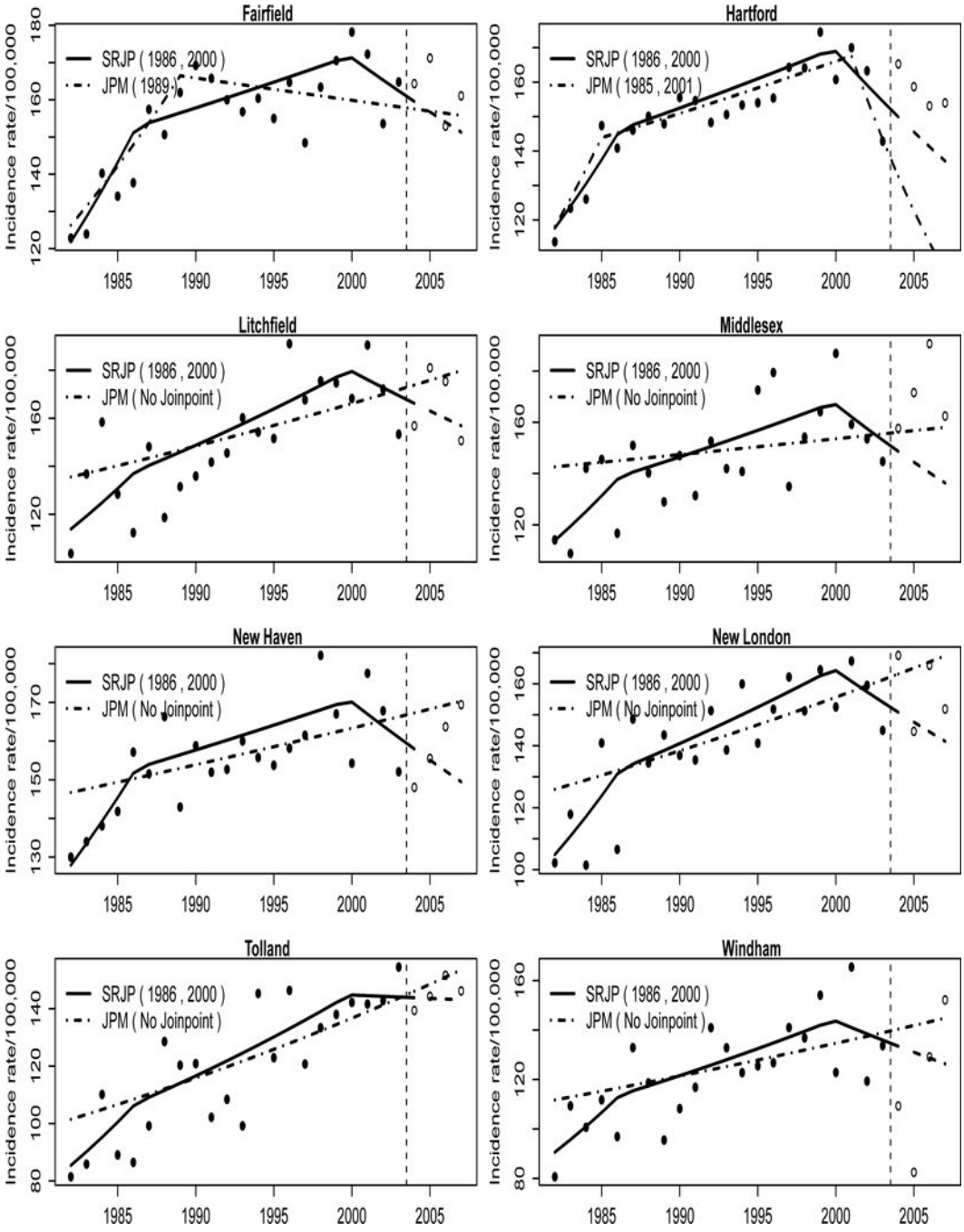

Figure 1.

Observed and fitted female breast cancer incidence rates (number of new cases per 100,000) for each Connecticut county

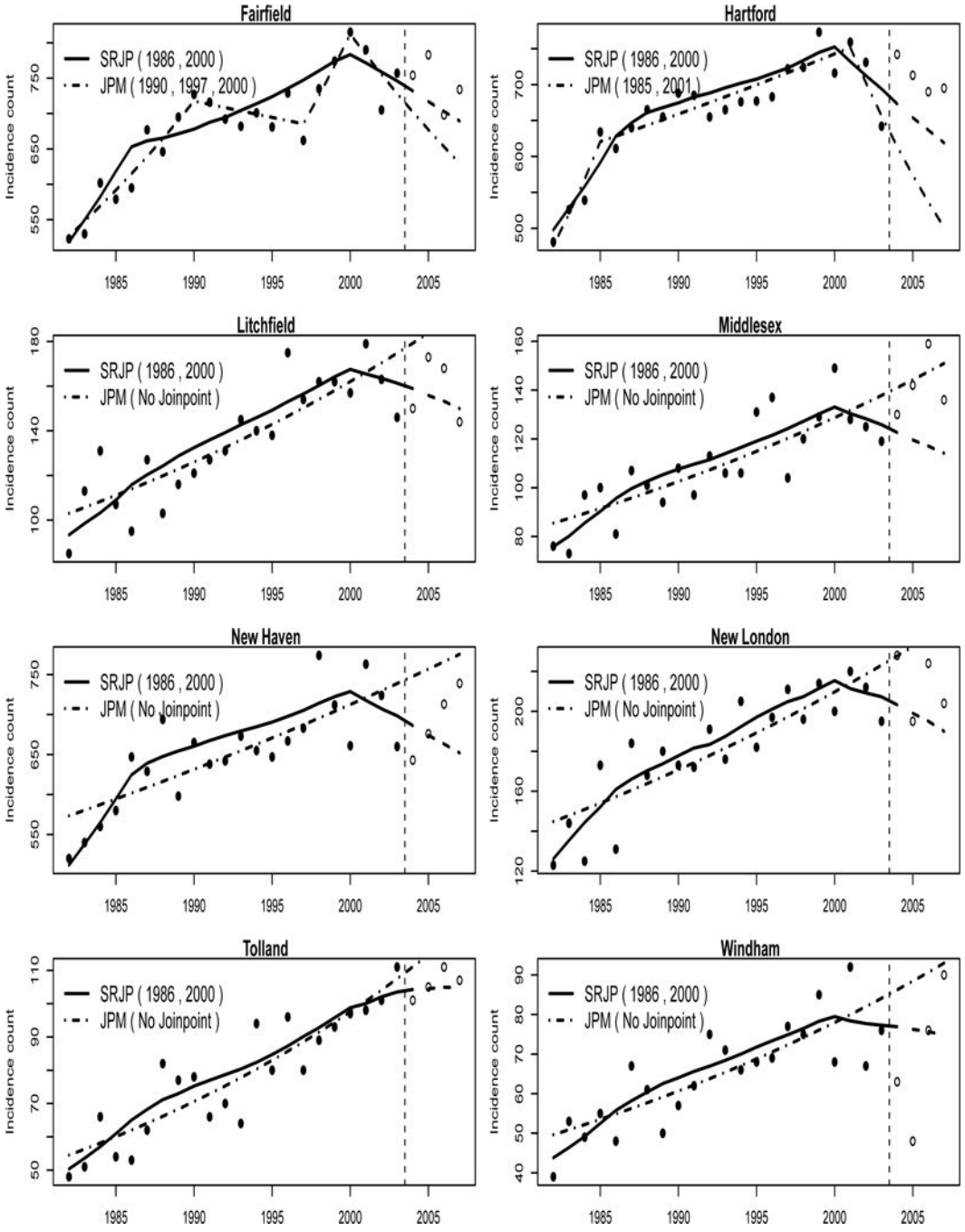

Figure 2.

Observed and fitted female breast cancer incidence counts (number of newly diagnosed cancer) for each Connecticut county

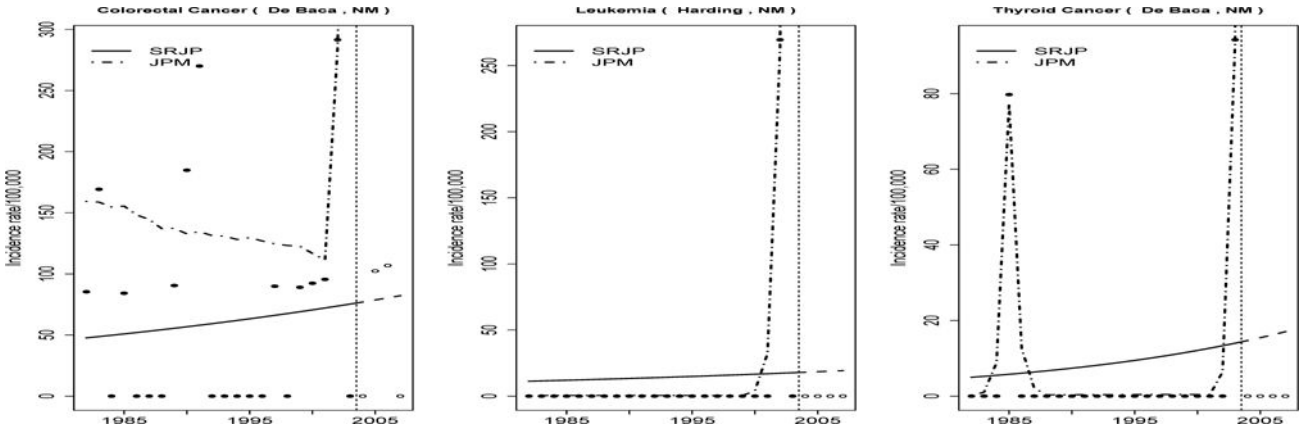

Figure 3.

Examples of three county-level incidence trends showing that the standard JPM may break down

Stat Med. Author manuscript; available in PMC 2018 May 03.

Table 1

Annual county-level cancer incidence rates per 100,000 and counts by cancer site for the counties under study (1982–2007)

	Men						Women					
	Incidence rate			Incidence count			Incidence rate			Incidence count		
	Mean	Median	Max	Mean	Median	Max	Mean	Median	Max	Mean	Median	Max
All sites	530	528	1361	134	48	2595	453	456	1535	126	42	2557
Female Breast	.	.	.	.	.	.	132	128	623	38	12	815
Cervix Uteri	.	.	.	.	.	.	9	6	304	3	1	61
Colorectal	67	62	411	16	6	353	67	60	321	10	6	158
Esophagus	7	0	270	2	0	52	2	0	94	1	0	22
Hodgkin’s Lymphoma	3	0	72	1	0	29	3	0	224	1	0	23
Kidney Renal Pelvis	16	12	223	4	1	99	10	7	262	3	1	71
Larynx	7	2	192	2	1	58	1	0	59	0	0	19
Leukemia	19	15	270	4	2	92	13	10	251	3	1	66
Lung/Bronchus	84	81	563	20	7	438	45	40	311	14	4	365
Melanoma	17	14	424	5	1	159	14	11	306	4	1	122
Myeloma	7	2	185	2	1	39	6	0	287	1	0	33
Non-Hodgkin’s Lymphoma	20	17	269	5	2	131	18	15	258	5	2	119
Oral cavity Pharynx	19	14	291	4	2	88	7	1	173	2	1	49
Ovary	.	.	.	.	.	.	17	14	439	5	2	95
Pancreas	13	10	262	3	1	76	12	9	260	3	1	82
Prostate	153	141	711	38	13	932	.	.	.	.	.	.
Stomach	10	6	213	3	1	77	5	0	299	2	0	50
Testis	5	0	279	2	0	48	.	.	.	.	.	.
Thyroid	4	0	219	1	0	52	10	7	249	3	1	158
Urinary Bladder	35	31	278	9	3	217	11	8	246	3	1	90

Table 2

Summary statistics of the PE measure for the prediction of incidences in 2004–2007 for Connecticut, Iowa, New Mexico and Utah

Outcome	Model	Mean	Median	p2.5%	p97.5%	Minimum	Maximum
Incidence rate	JPM	14.49	5.03	−57.48	136.39	−719.95	1437.75
Incidence rate	SRJP	−0.21	1.94	−72.47	62.95	−712.35	637.24
Incidence count	JPM	1.54	0.64	−6.236	14.20	−193.76	474.77
Incidence count	SRJP	0.37	0.16	−7.428	9.41	−173.87	288.31

Table 3

Medians of three measures of prediction errors of incidence rates per 100,000 for the years 2004–2007 for the 168 counties under study

	Men						Women					
	PE		APE		RPE (%)		PE		APE		RPE (%)	
Cancer site	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**
All sites combined	26.8	7.5	81.4	62.9	15	12	17.4	6.2	65.2	60.8	14	13
Prostate/Female Breast	27.3	−0.3	50.3	34.0	33	22	20.0	−3.0	36.5	27.5	28	21
Cervix Uteri	.	.	.	.	.	.	5.9	2.6	8.1	6.6	89	73
Colorectal	10.6	8.0	23.2	19.6	34	29	8.8	5.1	20.0	17.6	30	26
Esophagus	5.0	2.5	8.3	7.1	119	101	**1.0**	**1.7**	2.6	2.5	134	130
Hodgkin’s Lymphoma	**2.2**	**2.4**	4.4	2.8	148	94	**1.9**	**2.2**	3.5	2.7	135	103
Kidney Renal Pelvis	4.1	0.5	11.7	9.8	75	63	4.9	0.9	10.0	7.8	100	78
Larynx	3.8	1.7	6.4	6.0	89	83	**0.8**	**1.4**	**1.4**	**1.6**	**95**	**114**
Leukemia	7.6	3.3	11.7	9.3	61	49	5.1	2.0	9.8	8.4	73	63
Lung and Bronchus	8.6	4.0	23.5	19.7	28	23	8.0	1.9	18.1	16.2	41	36
Melanoma	7.6	−0.6	12.8	11.0	76	65	5.8	−0.7	11.9	10.7	84	75
Myeloma	5.4	3.7	9.5	7.0	142	105	5.7	3.4	7.6	5.2	136	92
Non-Hodgkin’s Lymphoma	5.4	2.1	14.0	11.5	69	57	6.3	1.8	12.3	10.2	67	56
Oral cavity Pharynx	3.8	−0.5	10.1	9.0	54	49	4.7	2.3	7.6	6.2	115	94
Testis/Ovary	4.6	3.9	6.8	5.1	135	100	6.3	2.5	10.4	9.3	62	56
Pancreas	5.1	0.5	9.4	7.4	75	59	5.5	1.4	9.4	7.8	80	65
Stomach	4.3	0.7	8.4	6.9	86	71	4.6	2.8	6.8	5.0	126	93
Thyroid	**2.2**	**3.2**	5.6	4.5	158	128	4.3	0.6	10.6	9.2	107	93
Urinary Bladder	6.0	1.8	13.7	13.4	39	38	5.4	2.6	9.1	6.9	86	66

Table 4

Medians of the measures of prediction errors of incidence counts (number of new cancer cases) for the years 2004–2007 for the 168 counties under study

	Men						Women					
	PE		APE		RPE (%)		PE		APE		RPE (%)	
Cancer site	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**	**JPM**	**SRJP**
All sites combined	4.0	0.8	7.7	6.3	6	5	4.1	0.6	6.8	5.8	5	5
Prostate/Female Breast	3.7	−0.1	5.5	3.3	15	9	2.7	−0.2	3.8	3.0	10	8
Cervix Uteri	.	.	.	.	.	.	0.7	0.3	0.8	0.5	33	21
Colorectal	1.4	0.8	2.5	1.9	15	12	1.0	0.5	2.2	1.7	13	10
Esophagus	0.6	0.2	0.8	0.7	45	38	0.5	0.1	0.6	0.2	98	37
Hodgkin’s Lymphoma	0.5	0.1	0.6	0.3	63	30	0.6	0.2	0.6	0.3	76	33
Kidney Renal Pelvis	0.6	0.1	1.3	1.0	33	25	0.6	0.1	1.0	0.7	39	29
Larynx	0.6	0.1	0.8	0.5	40	28	0.5	0.1	0.6	0.2	121	32
Leukemia	0.8	0.4	1.3	1.0	29	23	0.6	0.2	1.1	0.9	33	27
Lung and Bronchus	1.3	0.3	2.5	2.0	12	10	1.2	0.2	2.0	1.6	15	12
Melanoma	0.6	−0.1	1.5	1.1	28	20	0.6	−0.1	1.1	1.0	25	22
Myeloma	0.7	0.3	0.9	0.6	54	36	0.6	0.2	0.8	0.5	53	36
Non-Hodgkin’s Lymphoma	0.6	0.2	1.4	1.2	27	23	0.7	0.2	1.4	1.0	28	21
Oral cavity Pharynx	0.4	−0.1	1.1	0.9	25	21	0.7	0.2	0.8	0.6	43	28
Testis/Ovary	0.6	0.2	0.7	0.5	45	32	0.6	0.3	1.1	0.9	24	21
Pancreas	0.5	0.1	1.0	0.7	31	23	0.6	0.1	1.1	0.8	33	25
Stomach	0.5	0.1	0.8	0.6	30	23	0.6	0.2	0.7	0.5	42	29
Thyroid	0.5	0.2	0.7	0.5	60	44	0.5	0.1	1.0	0.8	31	25
Urinary Bladder	0.7	0.2	1.6	1.3	17	15	0.6	0.2	1.0	0.7	31	22

Table 5

Predicted prostate cancer incidence rates and counts in 2011 for the eight counties in Connecticut

County	Population	Rate	(95% CI)	Count	(95% CI)
Fairfield	451,000	193	(177, 211)	872	(798, 951)
Hartford	432,843	181	(166, 198)	785	(719, 857)
Litchfield	92,837	178	(157, 202)	166	(146, 187)
Middlesex	81,125	222	(195, 252)	180	(158, 205)
New Haven	414,838	164	(150, 180)	682	(622, 747)
New London	136,578	148	(132, 166)	202	(180, 227)
Tolland	77,006	157	(136, 180)	121	(105, 139)
Windham	58,589	101	(85, 120)	59	(50, 70)