Statistics and Its Interface Volume 7 (2014) 135–151

Developments and challenges in statistical methods in cancer surveillance

Huann-Sheng Chen*, Angela B. Mariotto, Li Zhu,
Hyune-Ju Kim, Hyunsoon Cho, and Eric J. Feuer

Cancer surveillance includes the monitoring of population levels and trends in incidence, survival, mortality, and prevalence. In addition, data are collected on the factors that influence these basic statistics across the entire cancer control continuum, such as healthy populations at risk of cancer, new diagnosis of cancer, treatment of cancer, living with cancer, and dying of cancer or other causes. To interpret the cancer statistics that are collected, an entire area of statistical methodology has been developed at the U.S. National Cancer Institute (NCI) and other institutions throughout the world. Most of these developments took place in the last 20 years, and the field is still evolving. In this review, we provide an overview of these methods, including the motivation for their development and how the methods compare with more general mainstream statistical methodology; available software; and relevant literature references.

AMS 2010 subject classifications: 62P10.
Keywords and phrases: Cancer surveillance, Joinpoint regression model, Delay adjustment model, Survival analysis, Spatial statistics, Incidence, Mortality, Prevalence.

1. INTRODUCTION

Cancer surveillance is the final phase of cancer research. The first phase includes basic research in cancer biology, genomics, imaging, biomarkers, and other areas. In the second phase, basic research discoveries are translated into specific technologies (e.g., a new screening test) that can be applied in the population. Cancer control, the third phase, is the scientific study of mechanisms to gain full implementation of proven technologies. Cancer surveillance is the monitoring of population levels and trends in incidence, survival, mortality, and prevalence (the so-called "big four" cancer statistics).

Two of the "big four" statistics (incidence and survival) are estimated from data collected from population-based cancer registries, which in the U.S. include the National Cancer Institute's Surveillance Epidemiology and End Results Program (NCI-SEER) and the Centers for Disease Control and Prevention's National Program of Cancer Registries (CDC-NPCR). Together, these sets of cancer registries provide almost complete coverage of the U.S., and SEER has registries that cover a long period of time but not for the entire U.S. The North American Association of Central Cancer Registries (NAACCR) is a member organization to which all U.S. registries belong. In the U.S., mortality data are collected from state reporting agencies by the National Center for Health Statistics. The final "big four" statistic, prevalence, is derived from incidence, survival, and sometimes also mortality data, depending on the model used for estimation.

Trends in cancer mortality reflect the ultimate success or failure of all the prior steps of cancer research. However, for cancer surveillance to be successful, it should enable feedback to all of the earlier stages of cancer research to help the entire cancer research enterprise optimize its chance of having the largest population impact. To do this, data are collected beyond the "big four" cancer statistics, including information on those factors that influence these basic statistics across the cancer control continuum, i.e., healthy populations at risk of cancer, new diagnosis of cancer, cancer treatment, living with cancer, and dying of cancer or other causes (Figure 1) [1, 2]. These data elements (e.g., cancer screening rates, dissemination of new treatments, and risk factors in the population) often must be collected through nationally representative sample surveys.

To interpret the cancer statistics that are collected, an entire area of statistical methodology has been developed. Much of this methodology has been developed in the last 20 years, and the field is still evolving. While these methods share some similarities and a common purpose with more general statistical methodology, they have been tuned to the needs of the cancer surveillance community. Many of these methods have been developed in the NCI's Surveillance Research Program, although there have been extensive developments outside of NCI as well (especially in Europe). In addition to the temporal trends of the "big four" cancer statistics, the analysis and interpretation of geographical and spatial patterns of cancer trends are also critical. Figure 2 highlights five areas where statistical methods and associated software have been developed by NCI. The purpose of this review is to provide an overview of these meth-

*Corresponding author.

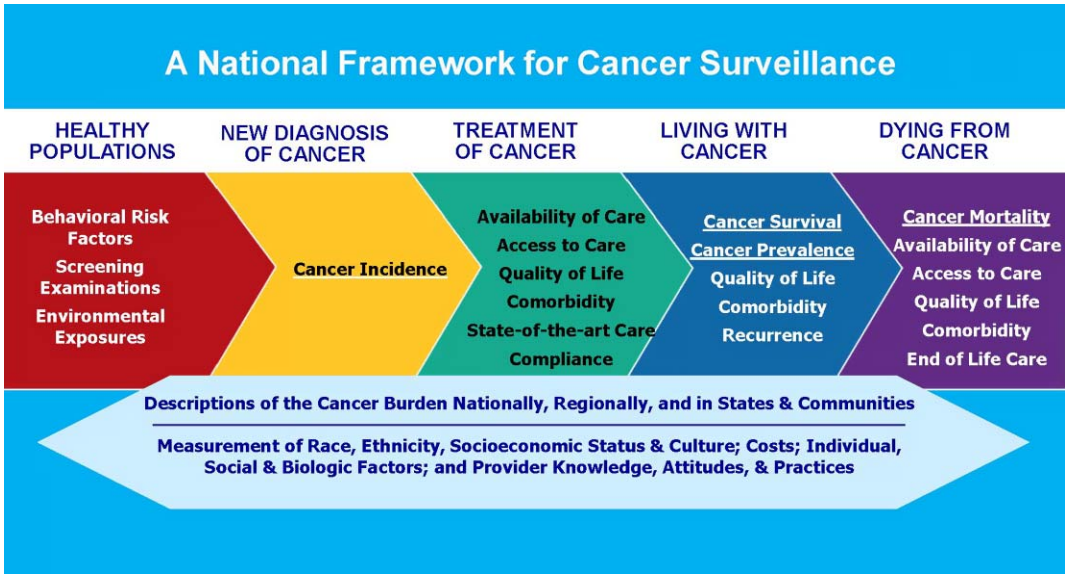

Figure 1. A flow chart of cancer surveillance.

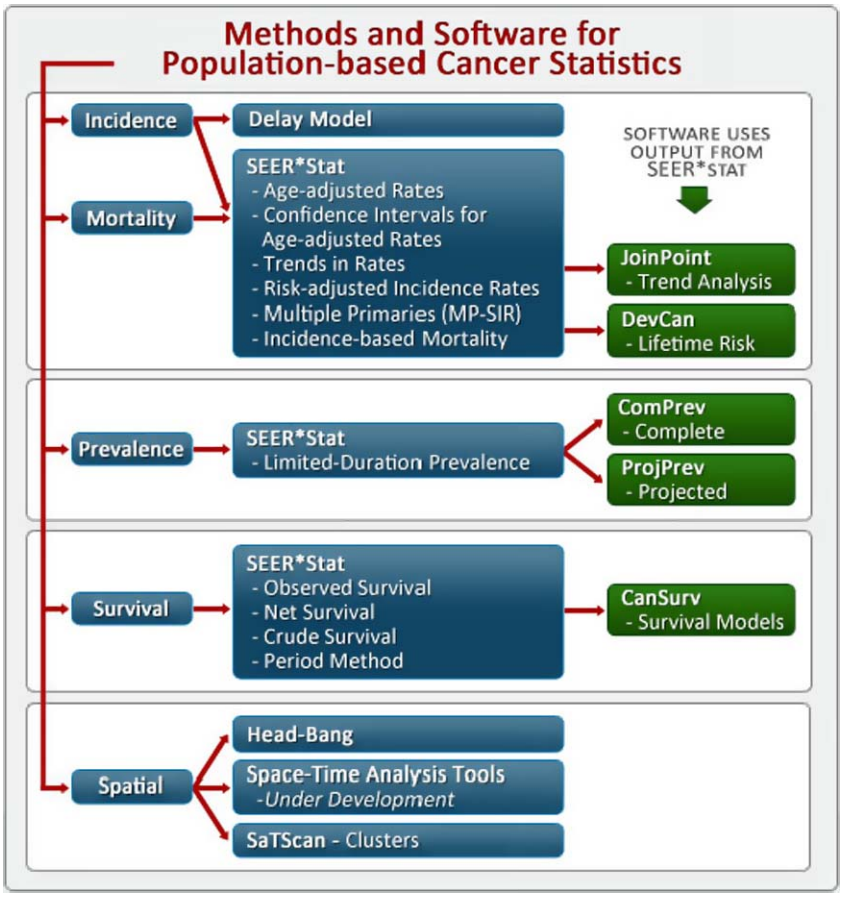

Figure 2. A diagram of methods and software developed by NCI for cancer surveillance.

ods, including the motivation for their development and an explanation of how the methods compare to more general mainstream statistical methodology; available software; and relevant literature references. Statistical details are not provided, and the interested reader should refer to the referenced articles for additional information.

2. ANALYSIS OF INCIDENCE AND MORTALITY TRENDS

2.1 Joinpoint modeling – is the trend changing?

The most frequently asked question in the analysis of cancer trends is, "Is the trend changing?" This seemingly simple question is surprisingly difficult to answer using standard statistical methods. In the past, various methods were employed, but these methods had distinct disadvantages. A polynomial fit to the data has a continuously changing slope, which makes it difficult to answer the question. In addition, polynomials sometimes fit poorly at the ends of the data. A test of the difference in slopes of a linear model fit to the log of age-adjusted rates over two fixed periods (e.g., the last five years and the five years prior to that), yields an annual percent change (APC) for each interval. However, the choice of the intervals is pre-specified rather than determined by the data, and the model assumes linearity (on a log scale) over each interval.

To describe changes in cancer trends, Kim et al. [3] used a segmented linear model and proposed a method to select the number of segments and estimate the model parameters. More formally, the following model is considered to describe cancer rates or more specifically, the log of rates (y) over time (x):

$$y = \beta_0 + \beta_1 x + \delta_1 (x - \tau_1)^+ + \cdots + \delta_\kappa (x - \tau_\kappa)^+ + error,$$

where $a^+ = \max(a, 0)$, the τ's are unknown locations of change-points where segment mean functions change, and the number of change-points, κ, is assumed to be unknown. Such a segmented line regression model in which linear segments are assumed to be continuous is called a joinpoint regression model, and the points where the regression mean functions change are called joinpoints [3]. The least squares method was used to fit the model with a given number of joinpoints, $\kappa = k$. To estimate the locations of k unknown joinpoints, $\tau_1, \cdots, \tau_k$, the grid search method described by Lerman [4] was used. Later, a continuous fitting method proposed by Hudson [5] was implemented into the model [6]. Hudson's approach allows the estimated joinpoints to be anywhere in the data range. The overall least squares estimates of the regression coefficients are then obtained based on the estimated joinpoints. Once the least squares fit is obtained for a model with $\kappa = k$, it is tested to determine whether addition of joinpoints significantly reduces the residual sum of squares. Starting with testing the null hypothesis that there are k_0 joinpoints against the alternative hypothesis that there are k_1 joinpoints where $k_1 > k_0$, tests are repeatedly conducted until for some k, the testing of $H_0 : \kappa = k$ versus $H_1 : \kappa = k+1$ is performed. Motivated by the fact that classical asymptotic theory does not work in this situation, a Monte Carlo method is used to estimate the p-value of each test. Because the procedure is based on multiple testing, the significance level of each test is adjusted to maintain the overall level under α, which is the probability of over-fitting the model.

Since Joinpoint software, which implements the method proposed in Kim et al. [3], was first released in 1998, a number of improvements have been made (http://surveillance.cancer.gov/joinpoint/). For example, the computational efficiency of the permutation procedure, which requires a lengthy computation to resample data points, was improved by using a sequential stopping rule. The main idea of sequential stopping is to stop resampling if the early replications of data indicate the p-value with the entire resampling to be large or small enough to draw a conclusion. Sequential stopping methods based on a truncated sequential probability test [7] as well as a simple curtailed test were implemented to perform the permutation test more efficiently. Model selection methods based on the Bayes Information Criterion (BIC) and a modified BIC, which work as faster alternatives to the permutation procedure, were also included. Detailed comparisons of these model selection procedures and user guidelines on which model selection method to use will be provided in a future paper.

After the number of segments is determined, asymptotic inference on regression parameters, including slope parameters and joinpoint locations, is performed. The p-values and confidence intervals for the regression slopes or equivalently for the APC rates are based on asymptotic normality, originally proved by Feder [8] and later by others ([9], [10]). Due to a slow convergence of the distribution of the estimated joinpoint to a normal distribution, however, a likelihood method was used to construct confidence intervals for the joinpoint locations. Kim et al. [11] conducted extensive simulations to examine the accuracy of asymptotic inference under various conditions and validated empirical recommendations to omit data points that coincide with the estimated joinpoints and to use standard errors estimated without continuity constraints.

When reporting and comparing a large number of different cancer types in tabular form, it used to be common practice to present the APC and statistical significance of the final joinpoint segment. However, this approach was problematic because these segments all have different lengths, and the power of the statistical test that determines whether the APC differs from zero is dependent on the length of the segment. When comparing the final segment for two different cancers (with approximately equivalent underlying variability over time), one could have a final segment increasing at 1% per year for many years that is statistically significant, while the other could have a relatively short segment increasing at 4% per year that is not significant. A measure called average annual percent change (AAPC) was developed so that comparable recent trends of two or more series could be compared. The AAPC, which is estimated as a weighted geometric mean of the APC's over a fixed pre-specified seg-

ment, was developed by Clegg et al. [12]. The Annual Report to the Nation on the Status of Cancer, an annual publication summarizing trends in cancer rates, started reporting five-year AAPC's in 2009 [13]. Simulation work has shown that the original derivation of the AAPC confidence interval, which is based on asymptotic normality conditional on the estimated joinpoints, is generally quite conservative, and work on improvements is ongoing.

Another enhancement to the Joinpoint software was the addition of features that enable comparison of different groups (e.g., males and females) when trends for each group are modeled as a joinpoint regression model. Kim et al. [14] considered the problem of comparing two groups of trend data with similar characteristics. A permutation test was proposed to compare two segmented line regression functions to test whether two mean functions are (i) identical or (ii) parallel. The two-group comparability test is being extended to a multi-group situation in which the goal is to cluster groups with similar characteristics. The multi-group problem has two distinct sub-problems: (i) ordered groups (e.g., age) that should be contiguously clustered and (ii) unordered groups. The ordered group problem is especially important when decomposing trends in age-adjusted rates into clustered sub-groups. To cluster similar groups, either the permutation test or the BIC method is used to select the common number of joinpoints for each possible cluster and then search for the cluster breaking points that minimize the overall sum of squared errors. To determine the number of clusters, the BIC method is used based on our recent simulation study, which will be discussed in a future paper.

The Joinpoint model has been adopted by registries throughout the world to characterize population-based trends in cancer rates, and its use in characterizing other health indices is growing. The original Kim et al. paper [3] has been cited more than 800 times. MJ Schell [15] used the Joinpoint model to project the number of applied papers that cite a source paper 20 years after publication and used this approach to rank the top 50 applied statistics papers published between 1985 and 2003. The Kim et al. paper [3] was ranked 40th on that list.

2.2 Modeling reporting delays in cancer incidence

NCI-SEER collects cancer incidence data across U.S. registries. The registries submit their reporting of new diagnoses to the SEER Program, normally within a 22-month window. However, some newly diagnosed cancer cases may be reported after the 22-month required period. The discrepancy between the time of diagnosis and the first time reporting to NCI, called delay time, can lead to underreporting if not accounted for. Additionally, reporting corrections may occur during the reporting process. For example, recording of race may be corrected in subsequent years, or whether a cancer is primary or metastatic may be determined at a later date after first reporting. As records from various health care facilities are consolidated, it takes some time to ensure that they are from the same person and refer to the same cancer. The cases in a given diagnosis year are updated annually in subsequent data submissions. Updates include adding new but previously unreported cases, as well as deletion of existing cases due to corrections of race, cancer site, sex, and age at diagnosis. Other reasons that contribute to delay time and error include systematic changes of registry operations and sporadic and unpredictable changes from certain facilities that report cases, among others. All of these factors make monitoring cancer trends more difficult. In particular, interpretation of recent cancer trend changes becomes less reliable.

The statistical problem of delay adjustment is how to predict the true underlying cancer incidence given the incomplete data. The delay-adjustment problem in cancer statistics is similar to those encountered in many other applications, including HIV/AIDS studies in which time from HIV infection to AIDS must be predicted or AIDS must be predicted from reported cases [16–18] and in other contexts [19, 20]. However, the reporting corrections caused by reclassifying a reported cancer site to another site is unique to the cancer surveillance field and requires special attention.

To address the delay time problem, NCI adopted the delay adjustment model of Midthune [21] in 2003. Application of this model to estimate cancer incidence rates for nine SEER registries is described in Clegg [22]. The model considers both the observed added cases that were first reported to SEER and the dropped cases that were reported earlier and dropped due to correction. The added cases were assumed to follow a Poisson distribution, and the dropped cases were assumed to have a conditional binomial distribution given the past history of adds and drops. Specifically, if a_j is the added case reported at time j, d_j is the dropped case reported at time j, and n_j is the net count at time j, then

$$a_j \sim Poisson\,(\lambda p_j)$$

and

$$\begin{aligned}(d_j \mid a_k, d_k, k = 1, \ldots, j-1) &\sim (d_j \mid n_{j-1}) \\ &\sim Binomial\,(n_{j-1}, g_j)\,,\end{aligned}$$

where λ is the expected number of cancers eventually reported, p_j is the probability that a case is reported at delay time j, and g_j is the conditional probability of a reported case being removed at delay time j given that the case was reported and not removed before delay time j.

There are several options for parameterization of p_j. A simple model assumes a geometric distribution for p_j such that $p_j \sim \rho\,(1-\rho)^{j-1}$, $j = 1, \ldots, \infty$. However, this may

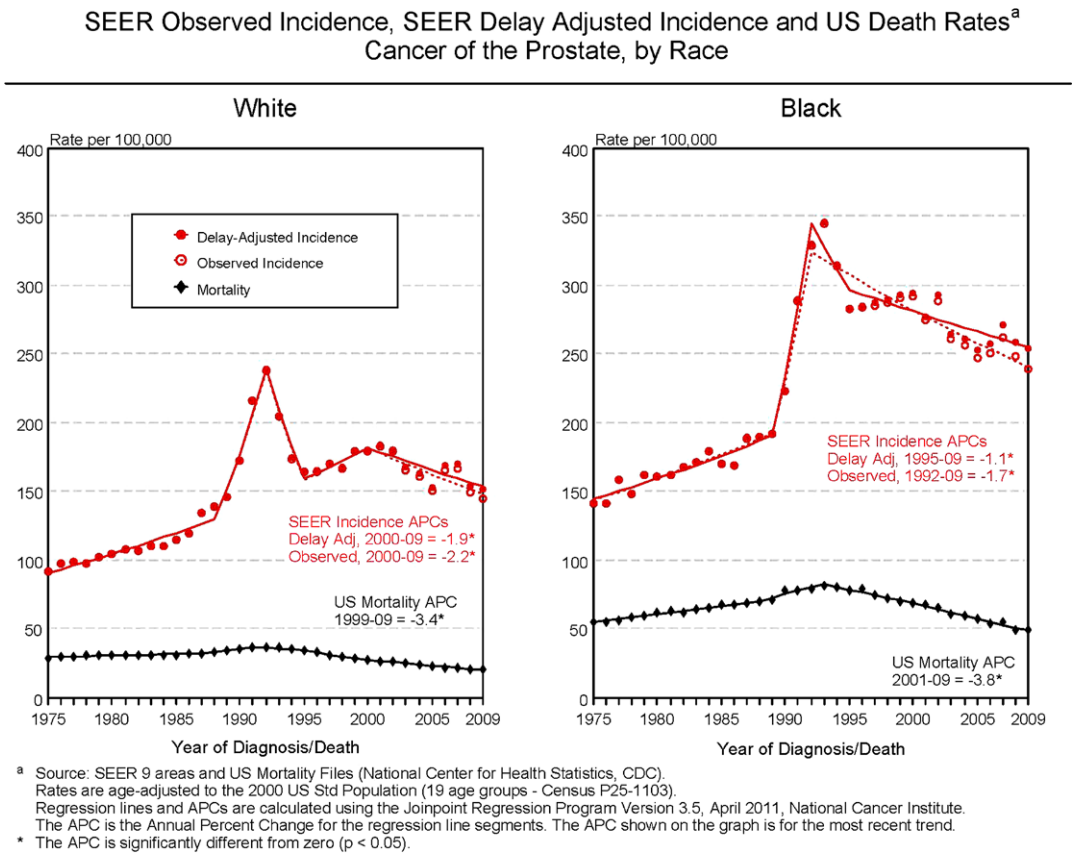

Figure 3. Joinpoint models for delay-adjusted incidence and observed incidence data of prostate cancer by race. Graphs were adapted from the SEER Cancer Statistics Review [30].

be impractical, assuming that the cases diagnosed in a certain year continued to be added thereafter infinitely, albeit rather slowly. A truncated delay distribution assumes that adding can only occur up to the longest reporting delay time among all the subpopulations considered. For the SEER 9 registries, this truncation point is not a problem because this set of registries has over 20 years of maximum reporting delay time, after which it is reasonable to assume that very few changes are made to the data.

Several factors can contribute to the variability of delay time distribution. These factors include diagnosis year, race, reporting resource, and registry. The delay time distribution is thus modeled as a function of these factors, either as a covariate or by stratification, depending on whether the proportional hazard assumption holds. Figure 3 shows an example of delay-adjusted prostate cancer incidence data for blacks and whites. The observed and delay-adjusted incidence rates of prostate cancer are shown in both panels. The dots are modeled by joinpoint regression models. The right panel shows that for blacks, the incidence rate is decreasing for both observed and delay-adjusted data, although delay-adjusted data decrease at a slower rate. The APC is -1.1 for the delay-adjusted rate and -1.7 for the observed data from 1995 to 2009. It also indicates that a trend change occurred around 1986 and then around 1992. The rapid rise and fall of the incidence trends are associated with the dissemination of new screening technology, in this case, the prostate-specific antigen (PSA) screening [23, 24].

The SEER 9 registries were expanded to SEER 13 in 1992 with the addition of the Los Angeles, San Jose-Monterey, Rural Georgia, and Alaska Native tumor registries. SEER 9 registries cover approximately 10% of the U.S. population, while SEER 13 registries cover approximately 17% of the U.S. population. SEER 9 has archived submission data back to diagnosis year 1981, which is 11 years more than the four registries added in 1992. This presented a problem because the original SEER 9 registries and the four new registries have different truncation points, making them difficult to compare. Statistically, the main concern is that the delay model assumes a truncated distribution, i.e., the probability of adding a new case is zero when the delay time is beyond the maximum length of the reporting years. To address this issue, we assumed that the data has two sets of maximum duration such that $J_1 > J_2$. For the data with shorter duration fitted by the delay-adjusted model, the probability of adding new cases (or percent missing) is zero when $t > J_2$, while data with longer duration fitted by its own delay-adjusted model can still find new cases when $t > J_2$.

To remedy this, Huang et al. assumed that after J_2 years, the percentage missing in the four expansion registries is the same as that in the original SEER 9 registries. Technical details are described in [25].

The problem of developing comparable delay adjustment factors across registries became even more complicated when SEER added four additional registries starting in 2000, and CDC-NPCR starting working to bring registries from the remainder of the country up to reporting standards, with some registries reporting cases diagnosed in 1999. With a large range of starting dates across registries, different registries having missing reporting years if their data did not meet certain quality standards, and different groups of registries potentially having different reporting delays, the modeling approach that had been applied to the expansion from SEER 9 to SEER 13 registries would be difficult to apply to registries across the U.S. A coordinated effort by NCI, CDC, and NAACCR is now under way to produce comparable delay-adjusted incidence rates for registries across the entire U.S. The approach to be employed will use data that have been reported to NAACCR for cases diagnosed as early as 1997. Groups of registries will be modeled together with those registries starting in later years or missing specific years borrowing information from registries that have complete data. The goals of this joint effort will be as follows: (i) Delay factors (and standard errors) for every (or almost every) U.S. registry will be produced; (ii) These factors should be easily combinable across registries so that analysts can obtain delay-adjusted incidence rates using any combination of registries; (iii) All delay adjustment factors across registries should be adjusted to the same truncation point; and (iv) Every case should have a delay factor attached to it. Ideally, computations for delay adjustment will be added to SEER*Stat (see Figure 2).

2.3 Incidence-based mortality (IBM)

Although U.S. mortality data derived from death certificates are often viewed as the ultimate indicator of cancer progress (because mortality is less influenced by biases than survival and incidence), these data lack information pertaining to the onset of disease, such as year of diagnosis, age at diagnosis, stage of disease at diagnosis, and histology of the tumor. For example, esophageal cancer has two major subtypes based on the histology of the tumor (adenocarcinoma and squamous cell carcinoma) with very different etiologies and population trends. It is not possible to estimate mortality trends for either of these subtypes of disease using U.S. mortality data because the histology of a cancer is not recorded on the death certificate. However, population-based cancer registries collect these types of data and allow the calculation of an incidence-based mortality rate [26]. This IBM rate allows a partitioning of mortality by variables associated with the cancer onset. IBM requires high-quality population-based cancer registry data and high-quality follow-up of cancer patients for vital status including cause of death.

If d_{ijk} is the number of deaths for age group i, calendar year j, and group k (e.g., histology), n_{ij} is the associated population, and w_i is the weight associated with a standard population, then the age-adjusted estimate of incidence-based mortality is

$$IBM_{jk} = \frac{\sum_i w_i \frac{d_{ijk}}{n_{ij}}}{\sum_i w_i}.$$

The sum of IBM across all levels of the factor k is the total estimate of incidence-based mortality, i.e.,

$$IBM_{j,Total} = \frac{\sum_i w_i \frac{\sum_k d_{ijk}}{n_{ij}}}{\sum_i w_i} = \sum_k IBM_{jk}$$

$IBM_{j,Total}$ may approximate death certificate mortality (DCM_j) but will not equal it exactly. Death certificate mortality comes from death certificates collected by states and represents all of the deaths which occur for residents of the state, regardless of where the person lived when they were diagnosed with the disease. On the other hand, Incidence-based mortality represents death certificates for everyone who was a resident of a registry catchment area (usually a state) when they were diagnosed with the cancer regardless of where they lived when they died.

If a registry reports cases from calendar year y onward (y $\leq j$), then $IBM_{j,Total}$ only includes deaths that occurred from cases diagnosed in year y or after, while DCM_j includes all deaths regardless of when they were diagnosed. Thus, one must allow a "burn-in" period during which the hazard of death from cancer becomes sufficiently small so that very few cases that were diagnosed before year y die in year j. For example, most people diagnosed with esophageal cancer die within three to four years of diagnosis. In this case, even though the cancer registries started in 1975, incidence based-mortality can only be estimated in 1978 and after. Secondly, there are differences between cases eligible to be captured by a cancer registry and the death certificates in a particular geographic area. Cancer registries capture every case diagnosed in their catchment area and follow them to death no matter where they live or when they die. Death certificates from the same catchment area are collected if someone dies in the area, regardless of where they lived or when they were diagnosed. As a practical matter, these biases (the out-migration after diagnosis and the in-migration prior to death) are relatively small, although in most cases DC mortality will be slightly larger than IB Mortality [26]. IBM can be estimated using SEER*Stat [27]. In analyzing IBM, one must be cautious in interpreting the results, because factors such as lead-time bias can influence these analyses (especially when partitioning IBM by year of diagnosis), whereas they generally will not influence overall

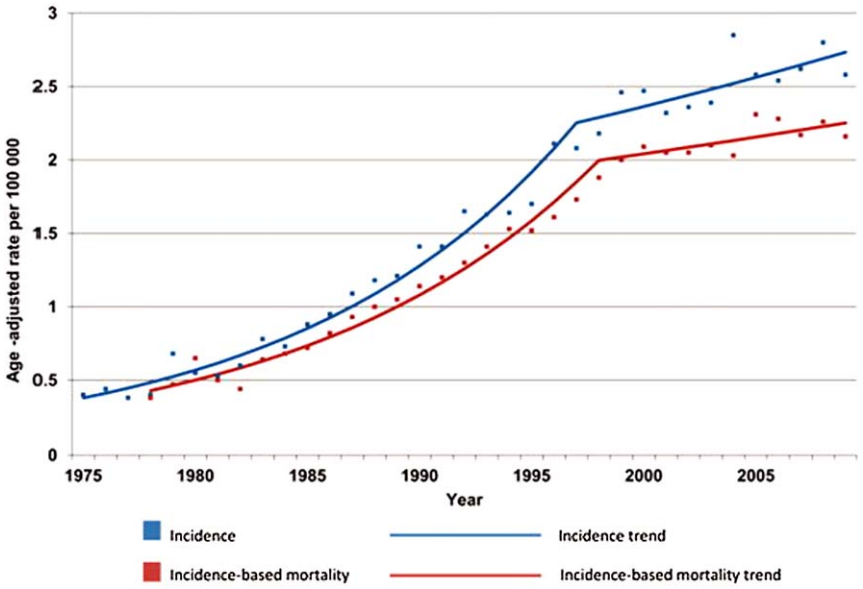

Figure 4. SEER 9 esophageal adenocarcinoma incidence and incidence-based mortality, 1975 to 2009. From 1975 to 1997, EAC incidence increased at an annual percentage change (APC) of 8.4 (95% confidence interval [CI] = 7.7–9.1), whereas the APC was 1.6 (95% CI = 0.0–3.3) from 1997 to 2009. For incidence-based mortality, the APC was 8.0 from 1978 to 1998 (95% CI = 7.2–8.8) and 1.1 from 1998 to 2009 (95% CI = −0.7 to 2.9). All rates were age-adjusted to the 2000 Standard population using 19 age groups. The graph was adapted from Figure 2 of Hur et al. [29].

death certificate mortality. Examples of IBM analyses include those for breast cancer [26], prostate cancer [28], and esophageal cancer [29]. Figure 4 shows trends in incidence and mortality for esophageal adenocarcinoma. Mortality is not shown for the first three years to allow for the burn-in period for IBM. A minimum three year burn-in period was considered sufficient in this case because three to four-year survival is so poor that very few cases diagnosed prior to 1975 would die of esophageal adenocarcinoma in 1978 or after.

3. SURVIVAL

3.1 Relative survival: not relying on cause of death information

The most frequently reported population-based survival statistic is relative survival as defined below [30], which is the measure used in comparisons of population-based cancer survival [31]. For most cancer registries, cause of death information obtained from death certificates is either unavailable or unreliable due to misclassification errors or intrinsic difficulties in identifying an underlying cause of death. For example, a metastasis site, rather than the original site of disease, might be reported as the cause of death. Relative survival was developed to estimate survival associated with a cancer diagnosis and does not rely on cause of death information. Relative survival $R(t)$ is defined as the ratio of observed survival (all-cause survival) of a cohort of cancer patients $S(t)$ to the expected survival of a comparable set of cancer-free individuals $S^*(t)$,

$$R(t) = \frac{S(t)}{S^*(t)}.$$

Thus, on the hazard scale, the overall hazard $h(t)$ is given by

$$h(t) = h^*(t) + \lambda(t),$$

where $h^*(t)$ is the expected mortality hazard and $\lambda(t)$ is the excess hazard associated with the disease of interest. Because a cohort of cancer-free individuals is difficult to obtain, life tables representing the survival of the general population are used to estimate expected survival/mortality. The underlying assumption is that the cancer deaths are a negligible proportion of all deaths [32]. Relative survival is usually estimated by using the actuarial method and dividing the time scale into intervals. Expected survival is calculated by averaging survival probabilities by age, sex, time period, and race from national life tables to individuals in the study population. Different methods have been developed (Ederer I [32], Ederer II [33], and Hakulinen [34]) to estimate expected survival. These methods differ regarding how long individuals from the study population are considered to be at risk to be matched and to enter expected survival calculations. A technical report [35] provides comparison of relative survival calculations by different methods. In practice, especially if the analysis is stratified by age, when estimating

short-term relative survival or calculating age-adjusted relative survival, the different methods do not make much difference and provide similar relative survival estimates [35]. However, in some particular situations, for example, for cancer sites diagnosed over a wide range of ages (e.g., thyroid), long-term relative survival for all ages combined may vary depending on the method used to estimate expected survival (see [35]). The Ederer II method has been shown to align well with the concept of net cancer survival [36] and is the default for calculation of relative survival in SEER*Stat. The standard error of relative survival can be estimated as the standard error of observed survival divided by the expected survival rate [32]. The standard error of observed survival can be estimated by Greenwood's formula [37].

3.1.1 Regression models for relative survival

The vast majority of work on survival modeling relates to cause-specific survival, i.e., survival calculated using cause of death. Recently, several methods have been developed and adapted to relative survival. Relative survival models may be classified into two broad groups: regression relative survival without cure and relative survival models with cure. Regression relative survival models (without cure) are used to estimate the effect of covariates on survival but also to project survival on both calendar time and follow-up time. Cure survival models aim to estimate the proportion of individuals who will eventually be cured and will not die of their cancer.

Most relative survival regression methods model the excess hazard $\lambda(t)$ of a cancer diagnosis [31, 38, 39]. Hakulinen and Tekanen [31] extended the grouped survival binomial regression model using a complementary log-log link to grouped relative survival data. Methods to estimate relative survival using individual data and the full likelihood approach have been developed by Dickman et al. [38] and Esteve et al. [39]. More recently, flexible parametric models [40, 41] have been developed to model relative survival by fitting restricted cubic splines on the log cumulative excess hazard scale. The main advantages of these models are the ability to model time on a continuous scale and the possibility to incorporate time-varying covariates. Other methods have been recently developed to model relative survival with time-dependent effects [40, 42–44].

3.1.2 Providing up-to-date estimates of survival

Survival data lag behind the current calendar year for two reasons: (i) the most recent year of diagnosis is usually three to four years behind the calendar year and (ii) typically only one year of follow-up information is available for patients diagnosed in the most recently reported year. Thus, five-year survival is usually estimated for patients diagnosed eight to nine years before the calendar year. However, there is interest in estimates that reflect the survival experience of the most recently diagnosed patients. The period method is a non-parametric method similar to cross-sectional life tables that uses the most recent follow-up information for each cancer patient to estimate survival [45].

In addition, regression models [31] that allow extrapolation of survival to the current calendar year and projection of survival into the future [46] have been used to model calendar year at diagnosis. Survival can be projected in two ways: (i) a flat projection extrapolates the fitted estimate of the last year of data to the current calendar year and (ii) a trend projection extrapolates the fitted trend to the current calendar year. Validation studies have shown that these models provide more accurate estimates of recently diagnosed patients than the period non-parametric method [45], [46].

3.1.3 Joinpoint survival analysis

Joinpoint models have recently been extended to model the progress of and trends in cancer survival rates [3], [47]. The survival joinpoint models fit linear segments to the hazard of dying and estimate changes in survival. In population-based surveillance, cancer survival trends are often characterized as a function of the year of diagnosis and projected into the current year. Under the proportional hazards assumption with joinpoints, the hazard function at t years after cancer diagnosis for a person who is diagnosed with cancer at calendar year x can be expressed as

$$(1)\quad \lambda(t \mid x, z) = \lambda_0(t) \exp\left\{\beta x + \sum_{k=1}^{K} \delta_k (x - \tau_k)^+ + \gamma^t z\right\},$$

where x is the calendar year of cancer diagnosis, z is the vector of covariates, $\tau = (\tau_1, \cdots, \tau_k)$ are the joinpoints, $\lambda_0(t)$ is the baseline hazard, and β, δ, γ are the parameters. Advances in early diagnosis and treatment often impact the survival of patients at a specific time point and then level off after those advances have been fully incorporated on a population level [48]. Therefore, Modeling survival trends and projecting up-to-date survival in the presence of a change point may facilitate understanding of the relationship between medical improvements and the survival experience for the patient population at large.

Yu et al. [47] extended the joinpoint survival model in equation (1) to population-based grouped survival data. They assumed that the number of deaths follows a binomial distribution; in relative survival analysis, the number of patients dying from all causes follows a binomial distribution. The joinpoint relative survival model can be fitted using SAS PROC GENMOD with a user-defined link function. An R package based on an iteratively reweighted least squares algorithm is currently being implemented and will be available to the public in the future.

A Bayesian approach to joinpoint survival models for population-based survival data has also been developed based on a Poisson distribution for the number of deaths and a Dirichlet process mixture for the regression slopes [49]. The Bayesian approach relaxes distributional assumption of

the regression coefficient by using a mixture of normal distributions. It can be applied to modeling rare cancers by using the Poisson distribution for the number of death. However, it is computationally more demanding with large datasets.

3.1.4 Cure survival models

Cure survival models for relative survival developed by Yu et al. [50] have been implemented in the CANSURV software. CANSURV software was developed to analyze grouped relative survival data, although some of the methods are available for individual data [50, 51]. CANSURV can fit parametric survival models with and without cure, and fit semi-parametric Cox proportional hazards method [50]. The CANSURV program implements parametric mixture cure survival models to analyze population-based cancer survival data:

$$S(t) = c + (1 - c)G(t \mid \mu, \sigma),$$

where c represents the proportion that will be cured and die of other causes, and $(1 - c)$ represents the proportion uncured whose cancer-specific survival follows the $G(t)$ distribution with mean μ and variance σ^2. In the CANSURV software, $G(t)$ can be a log-normal, log-logistic, or Weibull distribution. The model is described in detail in Yu et al. [50]. Recently, a cure survival model using a flexible parametric model for relative survival was developed [52]. The flexible parametric survival model use splines to model the underlying hazard, and therefore no parametric distribution has to be specified.

3.2 Cause-specific survival: improving cause of death information reported by cancer registries

The major advantage of relative survival (excess mortality) is that information on cause of death is not required. However, the method relies on accurately estimating expected survival using population life tables. If life tables are not representative of patients' survival in the absence of cancer then relative survival will be biased. For example, patients with smoking related cancers may experience excess mortality, compared to the general population, due to both the cancer and other smoking-related conditions. Thus their expected survival could be lower than that estimated from life tables and consequently, relative survival could be slightly overestimated [53]. On the other hand, relative survival of individuals diagnosed with localized breast or prostate cancer through screening, has shown to be higher than 100 percent [54], indicating that expected survival from life tables underestimate their survival due to other causes. These individuals probably go more frequently to clinicians indicating either a better access to care or healthier behavior than the general population (healthy screening effect). In addition, life tables are not available for all races and ethnicities, and the general population life tables may not be a good representation of their life expectancies. For these reasons NCI developed and improved an algorithm [54] to measure more accurately cause of death information. This algorithm considers causes of death that are likely to be related to the particular cancer or consequence of a cancer diagnosis and the fact that an individual may or may not have more than the particular cancer. In validation studies, the cause-specific survival using the new cause of death variable was similar to relative survival. It allows reporting of the most up-to-date cancer survival statistics on U.S. minorities such as Hispanics and Asians (e.g., Chinese, Japanese, Filipino, and Vietnamese) as well as Native Americans and Alaska Natives. More information of the cause of death algorithm can be found in [22, 54] and on the SEER website (http://seer.cancer.gov/causespecific/).

3.3 Competing risks

Cumulative mortality from cancer may be represented in the presence or absence of other causes of death using the theory of competing risks. Net measures of survival (survival in the absence of other causes of death) are more commonly reported from cancer registry data because this measure is useful in tracking the progress of cancer control efforts, since it is not influenced by changes in mortality from other causes. However only cumulative mortality in the presence of other causes is useful in estimating the actual survival patterns observed in a cohort of cancer patients.

Cronin and Feuer [55] developed a method for estimating crude cumulative mortality using a relative survival approach. Based on the theory of competing risks, for discrete time intervals, define the cumulative probabilities of death through time interval M from cancer, G_{cM}, and from other causes, G_{oM}, can be estimated as

$$\overline{G}_{cM} = \sum_{k=1}^{M} \left(\prod_{i=1}^{k-1} \hat{P}_i \right) \left[\left(1 - \overline{R}_k\right) - \frac{1}{2} \left(1 - \overline{R}_k\right) \left(1 - \overline{E}_k\right) \right]$$

$$\overline{G}_{oM} = \sum_{k=1}^{M} \left(\prod_{i=1}^{k-1} \hat{P}_i \right) \left[\left(1 - \overline{E}_k\right) - \frac{1}{2} \left(1 - \overline{R}_k\right) \left(1 - \overline{E}_k\right) \right]$$

where $\hat{P}_i$ is the life table estimate of the observed survival rate in interval i, $\overline{E}_i$ is the expected survival rate in interval i for the patient group as obtained from US Life Tables based on the age, race, sex, and calendar year of diagnosis mix of the group, and $\overline{R}_i = \frac{\hat{P}_i}{\overline{E}_i}$ is the estimated relative survival rate in interval i. This estimate assumes independence of the competing risks and that the conditional risk of dying from cancer and other causes follows a uniform distribution over each time interval. Figure 5 shows an example comparing net and crude cumulative mortality for localized prostate cancer for men age 70 and over. In the presence of other cause mortality (which is quite high for men 70 and over), the chance of death from prostate cancer is considerably reduced.

Feuer et al. [56] extended Cronin and Feuer [55] to the case of a specific cancer patient j with characteristics z_j

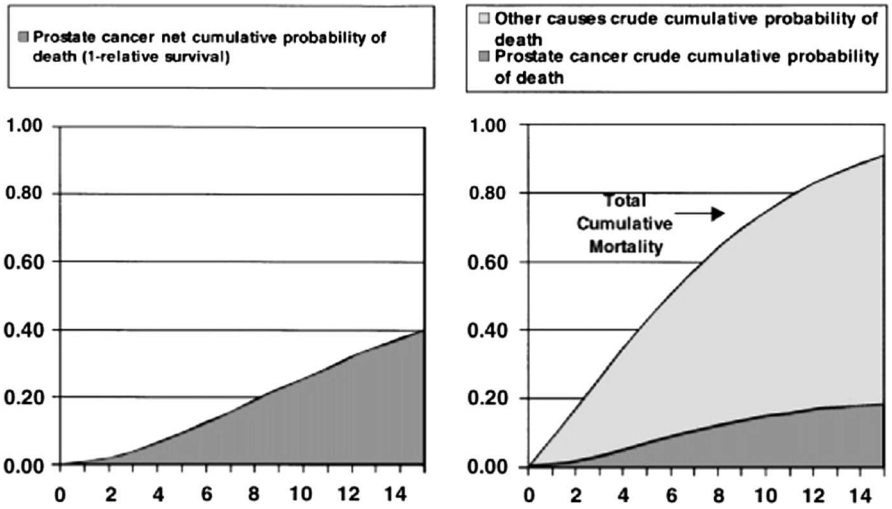

Figure 5. Cumulative probability of death in men with localized prostate cancer over the age of 70. The graphs were adapted from Figure 1 of Cronin and Feuer et al. [55].

which influences net cancer-specific survival, and characteristics w_j which influence net other cause survival [56, 57]. Generally z_j and w_j will have some factors in common (e.g. age). Generally cancer registries have many factors (z_j) available to characterize net cancer survival, but very few or no factors (w_j) besides age to characterize net other cause survival. Although independence of competing risks is a strong assumption, one advantage of operating under this assumption is that it is possible to estimate net other cause survival from alternative data sources other than the registry data.

Work is continuing at NCI on a Cancer Survival Query System [56], which will allow health care providers access to competing risks survival based on a large population-based cancer registry data base (i.e. SEER). If factors w_j characterizing other cause survival are not readily available in the cancer registry data, a potentially rich range of independent alternative data sources can be used. Of course, in using independent sources, one must assume that survival from other causes conditional on the available covariates is the same for the cancer and alternative data source. Based on SEER-Medicare linked data, Mariotto et al. [58] have developed estimates of co-morbidity based on claims in the year prior to diagnosis, but also included a non-cancer sample to help provide more stable estimates. Efforts are also underway to develop lifetables based on mortality follow-up from National Health Interview Surveys based on covariates such as self assessed overall health status, smoking, activities of daily living, etc.

Lee et al. [59] applied a more traditional approach developed by Cheng et al. [60] to cancer registry data which does not assume independence of competing risks. In this approach, cause of death is used and the chance of death from cancer and other causes are both estimated from a single data cancer registry data set.

4. CANCER PREVALENCE ESTIMATION

Prevalence is defined as the number or percentage of people alive on a certain date who were previously diagnosed with cancer. It includes new (incidence) and pre-existing cases and is a function of both past incidence and survival. Because it includes all prior diagnoses of cancer, prevalence is sometimes denoted "complete prevalence." "Limited-duration prevalence" refers to prevalence that includes survivors diagnosed a limited number of years prior to the prevalence date. Information on prevalence is crucial for health planning and resource allocation, and prevalence can serve as an estimate of cancer survivorship.

For the past 38 years, NCI has provided the nation's cancer prevalence estimates based on data from the SEER Program. However, the methods and data used to produce these estimates have evolved over time.

4.1 Limited-duration prevalence

The first national estimate of U.S. cancer prevalence used cancers diagnosed from 1935 through 1981 and follow-up through 1983 from the Connecticut registry [61]. This estimate represented 46-year limited-duration prevalence, i.e., people alive on January 1, 1982, who were diagnosed with cancer during the previous 46 years (1935–1981). This estimate did not represent complete prevalence, and because it was based on data from only the Connecticut tumor registry, it was not representative of the entire U.S. population. However, the method was adjusted for cases lost to follow-up by estimating their chance of being alive at the prevalence date and has been implemented in SEER*Stat software to estimate limited-duration prevalence. This method has also been used to estimate U.S. prevalence using data from the SEER 9 registries, which are more representative of the U.S. population.

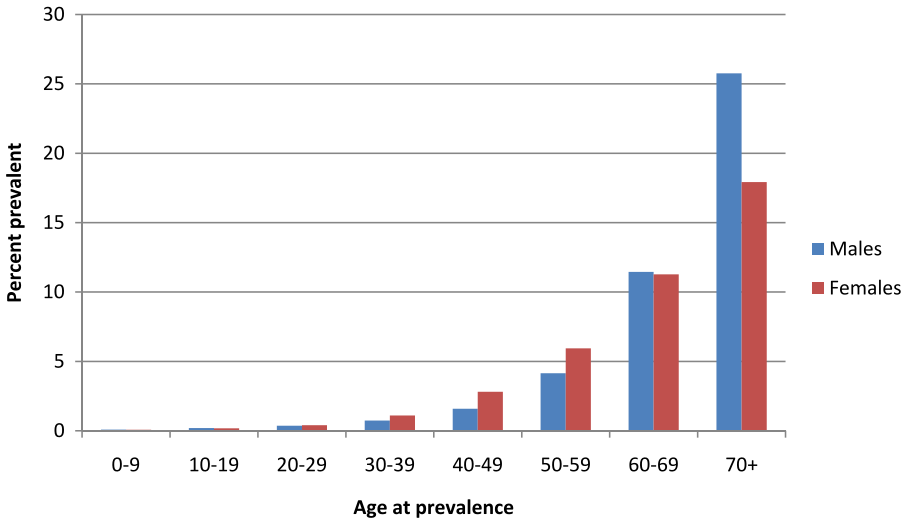

Figure 6. The percentage of the US population alive on January 1, 2009 with a prior diagnosis with cancer.

4.2 Complete prevalence

Complete prevalence is a more desirable statistic that includes survivors ever diagnosed with cancer. A method developed to estimate complete prevalence applies incidence and survival models [62, 63] to estimate the proportion of prevalent cases diagnosed prior to registration. This method has been used since 2002 to estimate complete prevalence in the U.S. and is implemented in COMPREV software [64]. Figure 6 shows the percentage of the U.S. population alive on January 1, 2009, with a prior diagnosis with cancer.

The complete prevalence at age x, $N(x)$, representing the proportion of individuals with cancer and aged x at the prevalence date, can be calculated as a convolution of incidence and survival as [65],

$$N(x) = \int_0^x I(t)S(t, x-t)dt \tag{2}$$

where $I(t)$ is the incidence hazard at age t and $S(t, x-t)$ the probability that individuals diagnosed with cancer at age t are still alive at age x. If we base our estimation on a registry that has been operating for l years, we can only estimate l years limited duration prevalence, i.e., the prevalence at age x of people diagnosed between ages $x-l$ and x, $\hat{N}_0(x; x-l, x)$. The COMPREV method [65], [66] uses the l-year limited duration prevalence $\hat{N}_0(x; x-l, x)$ and an adjustment factor based on estimates of modeled prevalence as specified in (2).

More specifically, complete prevalence is estimated by estimating parametric incidence and survival models, $\hat{I}$ and $\hat{S}$ and an adjustment factor based on the ratio of modeled complete prevalence and modeled observed prevalence, i.e.,

$$\begin{aligned}\hat{N}(x) &= \hat{N}_0(x; x-l, x)\frac{\hat{N}(x)}{\hat{N}(x; x-l, x)} \\ &= \hat{N}_0(x; x-l, x)\frac{\int_0^x \hat{I}(t)\hat{S}(x-t, t)dt}{\int_{x-l}^x \hat{I}(t)\hat{S}(x-t, t)dt}.\end{aligned}$$

Standard errors of a complete prevalence estimate can be calculated using the delta method [67]. This method has been extended to estimate complete prevalence for patients diagnosed during childhood (ages 0–19) [68, 69].

4.3 Projections of prevalence estimates

Usually, cancer registries do not cover the entire national population. In the U.S., the complete prevalence is estimated by extrapolation. Using SEER registry data, the prevalence proportion controlling for age, sex, and race (white, black and other races) is extrapolated to the respective national population [70].

However, two other relevant modeling methods exist: (i) the Prevalence-Incidence Aproach MODel (PIAMOD) based on equation (2) [71] and (ii) the Mortality-Incidence Approach MODel (MIAMOD) [72] method based on a back-calculation mortality equation. The MIAMOD method estimates prevalence from cancer mortality data, which is available nationally. The method models mortality and survival data to back-calculate incidence and then forward-calculate prevalence using equation (2). This method has been applied to estimate breast cancer prevalence at the state level in the U.S. [73].

Because prevalence is valuable in health planning and resource allocation, it is important to have future projections of prevalence. The PIAMOD method [74] is based on equation (2) and directly models incidence and survival to forward-calculate prevalence. This method has been used to project prevalence into the future and to investigate the

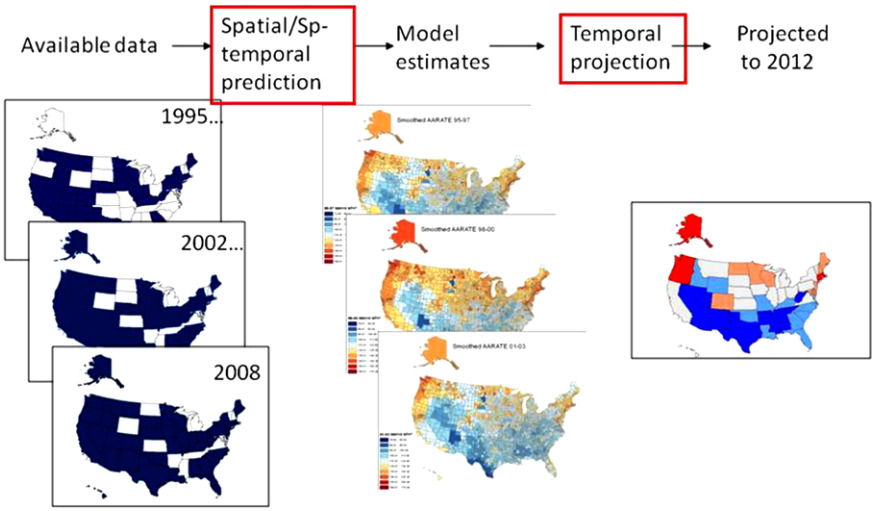

Figure 7. Overall process of the incidence spatial/temporal projection.

impact of changes in population, incidence, and survival in prevalence projections [71, 75]. It has also been used to estimate costs of care [71, 76].

5. SPATIO-TEMPORAL MODELING AND CLUSTER DETECTION

For more than half a century, the American Cancer Society has published the estimated cancer mortality and new cancer incidence in the current calendar year in the U.S. overall and in each state [77–79]. Cancer mortality and incidence cases diagnosed in the current calendar year in the U.S. overall and in each state are not known because the most recent year for which data are available lags three years (for mortality) or four years (for incidence) behind due to the time required for data collection, compilation, and dissemination. Furthermore, high quality incidence data have not yet been achieved in all states, and the total cases for the most recent one to three data years are incomplete because of delays in reporting. Until 2010, these estimates project three years (for mortality) [80] or four years (for incidence) [81] ahead of the most recent data year to the current calendar year.

However, in the past few years, because of delays in the release of the final mortality data for the most recent data year from the National Center for Health Statistics, it became necessary to develop a four-year-ahead projection method for mortality. The cancer mortality data are available over a long time period (since 1969) and cover all the states in the U.S.; in contrast, the cancer incidence data are available over a shorter time period (since 1995), do not fully cover all states in the U.S., and thus need a method to first fill in the "holes" in the data and then project four years ahead to the current calendar year. Beginning in 2010, a NCI work group was formed to evaluate projection methods for U.S.- and state-level cancer mortality and incidence. The details of the evaluation procedure and results were published in 2012 ([82, 83]).

In this section, a spatio-temporal projection method for incidence estimates will be described. The process used to estimate the numbers of new cancer cases expected in the current calendar year consists of three steps (see Figure 7):

I. Spatio-temporal prediction: A hierarchical Poisson mixed effects model [81] is applied to observed data from high quality cancer registries, as certified by NAACCR, to provide estimates of annual case counts over the available time period for every U.S. county. This step can fill in "holes" in a state's time series before the state became a certified high quality registry or fill in "holes" in the map for a year when some states did not report their data.
II. Delay adjustment: The predicted case counts from Step I are summed to the state level and then inflated to account for expected delay in case reporting.
III. Temporal projection: The delay-adjusted predicted case counts from Step II are projected ahead four years to the upcoming calendar year. For this validation test, the model projects ahead to the latest year for which observed data are available.

Because of the complexity of this process, validation of the spatio-temporal prediction and temporal projection steps are done separately. A residual analysis is performed on the results of Step I to determine whether additional covariates or interaction terms are needed. The temporal projection is validated by projecting delay-adjusted observed case counts four years ahead, comparing alternative methods by several fit statistics.

5.1 Data and methods for spatio-temporal predictions

An updated version of the CINA Deluxe incidence data from NAACCR described in Pickle et al. [81] was used in the spatio-temporal prediction validation. The data were compiled by NAACCR with the permission of its member registry and released to researchers for cancer surveillance research (see http://www.naaccr.org/Research/CINADeluxe.aspx). The updated dataset contains data from 1995 through 2007 and includes 46 states and the District of Columbia, covering 95% of the U.S. population [84].

The covariates for the spatio-temporal model are constructed from various sources. The only information available on the individual cases is age, gender, race, county of residence, cancer site, and year of diagnosis. Approximately 30 other ecologic covariates are available at the county level, including categories of socio-economic status available in the census data, availability of health services [85], behavior and risk factors [86], and the mortality data of the corresponding cancer site in the specific population [87]. A hierarchical Poisson regression model was used to estimate the number of cases for all U.S. counties by socio-economic and other ecological covariates. The number of new cancer cases in county i, age group j, and year t, denoted by y_{ijt}, was assumed to be distributed as a Poisson random variable, with mean $n_{ijt}\lambda_{ijt}$, where n_{ijt} is the corresponding population at risk and λ_{ijt} is the incidence rate. Assume a log-linear rate structure with

$$\log(\lambda_{ijt} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma, \boldsymbol{\delta}) = \alpha_r + f(a_j)\beta + \log(m_{ijt})\gamma + \mathbf{X}_i\boldsymbol{\delta} + \theta_s + \phi_t,$$

where α_r is the intercept for census region, a_j is the midpoint of a particular age group, $f(\cdot)$ is a cubic function of age groups that accounts for the non-linearity of some cancer rates among age groups, m_{ijt} is the age-specific mortality rate, and $\mathbf{X}_i$ is the vector of socio-demographic and lifestyle covariates. θ_s and ϕ_t account for spatial and temporal random effects, respectively. The spatial and temporal random effects would typically be specified in a fully Bayesian approach with conditional auto-regressive and auto-regressive priors, but the intensive computing and labor involved makes it infeasible to run the program for the many cancer sites that NCI projects. As a result, a low-rank smoothing (the number of smoothers is considerably less than the number of observations) in SAS Glimmix procedure [88] was used, which constructs the "knots" of both spatial and temporal random effects so that the knots are equally spaced and cover the whole U.S. for the study period.

5.2 Data and methods for temporal projections

To date, not enough incidence data are available from every registry to test a projection four years ahead for the entire country. The latest CINA Deluxe dataset now includes incidence data for 1995 through 2008, a 14-year span. Because one of the test methods (Nordpred) requires that time be specified in five-year blocks, the required observed data time span was extended to 15 years. Thus, 19 years of observed data are required for this projection: 15 years (1990–2004) for model input plus four years for projection ahead (to 2008). Only the older SEER 9 registries can provide sufficient data for this projection. In addition, the rest of California and the state of New Jersey have available data and gave permission for their use. The aggregate of the SEER 9 registries plus these two additional areas were used as a proxy for the entire U.S. The selection of cancer sites includes both very common and very rare cancers and is the same set of sites used for the original model development.

A more thorough search was conducted for a better temporal projection method to project four years ahead for cancer incidence counts, the third step of the three-step prediction process. In this search, five projection methods were evaluated and compared: the Nordpred method [89], (a special version of the Age-Period-Cohort method [90], Joinpoint method (the method previously used [81]), State-Space model which used a local quadratic function to obtain projections of the time series [80], Bayesian State-Space (BSS) method, and Vector Autoregressive (VAR) model. BSS used a dynamic generalized linear model fitted in the Bayesian paradigm. It first modeled the logarithm of the parameter (state) of the assumed Poisson distribution for the incidence counts for the initial year, and then combined the amount of variation from year to year to put together the likelihood. VAR applied the empirical mode decomposition [91] method to decompose the data and then applied multivariate time-series technique for projections. At the U.S. level, on average, the Bayesian State-Space method produced projections that are closer to the observed counts than the other projections. But for the most common cancer sites, Vector Autoregressive model outperforms the Bayesian State-Space method. At the state level, the Vector Autoregressive model produces projections that are closest to the observed counts. Although two methods could be recommended – the Bayesian State-Space method at the U.S. level and the Vector Autoregressive model at the state level – the decision was made to use a single model, the Vector Autoregressive model, at both the U.S. and the state levels starting with the Cancer Facts & Figures 2012 and Cancer Statistics 2012.

5.3 Cluster detection using scan statistics

Cancer registries and public health investigators may be interested in emerging spatial patterns and/or temporal trends of cancer rates when new cancer data become available every year. Cluster detection methods using spatial and space-time scan statistics have become very popular in recent years, largely due to the open-source software SaTScan [92]. In such methods, a variably shaped and sized candidate area (scan window) scans across a study region and

period. For each window, a likelihood ratio is calculated, and the window with the maximum likelihood ratio is selected as the most likely cluster. This is the cluster that is least likely to have occurred by chance. The likelihood ratio for this window constitutes the maximum likelihood ratio test statistic. Its distribution and p-value are obtained through the Monte Carlo hypothesis testing method. Several methods have been proposed in this field, including Kulldorff's scan statistics [92, 93], flexible-shaped [94], upper-level set [95], and other likelihood-based methods [96].

A purely spatial scan statistic imposes a circular or elliptic (with different shapes and angles) window over the study area in searching for clusters. Extended to the time dimension, a space-time scan statistic is defined by a cylindrical window with a circular or elliptic spatial base and height corresponding to time, which denotes the time period of potential clusters. In the equation below, Z represents the collection of all the possible clusters z in study region S. A zone (z) consists of neighboring geographic units having their centroids in circles (or ellipses) of various radii (and orientation). The variables c_z and n_z represent the observed number of cases and the expected number of cases (or population) in zone z, respectively. Thus, $C = \sum_z c_z$ and $N = \sum_z n_z$ represent the total number of cases and the total number of expected cases (or population) in S, respectively. For cancer incidence and mortality, a Poisson model is typically chosen. The likelihood ratio of a zone (z) is then given by

$$LR(z) = \left\{ \left(\frac{c_z}{n_z} \right)^{c_z} \left(\frac{C - c_z}{N - n_z} \right)^{C - c_z} \right\} I(c_z > n_z).$$

The most likely cluster is the scanning window $z \in Z$, which maximizes the log likelihood ratio. A closed form of the null distribution of the test statistics $T = \max_z \log(LR(z)) = \max_z LLR(z)$ does not exist, so statistical significance is evaluated using Monte Carlo hypothesis testing.

The SaTScan website (http://satscan.org) lists a large number of publications in the methodology and applications by field of study in cluster detection. Of special interest to cancer cluster detection are investigation into breast cancer in the northeastern U.S. [97], prostate cancer mortality in the U.S. [98], and brain cancer in the U.S. [99].

6. DISCUSSION

Cancer statistics derived from population-based cancer registries can be classified into two general categories. The first category includes statistics that are derived for policy purposes and are generally normalized or adjusted in some manner to control for confounding factors. For example, age-adjusted rates allow comparisons across years without the confounding effects of changes in the age distribution of the U.S. population. Trends in the five-year net relative survival over time reflect only the impact of changes in the hazard of cancer death without the confounding effects of changes in the hazard of death from other causes. The second category of statistics is developed to be more applicable to individuals or to show the population burden of disease. Competing risks estimates of survival are applicable to actual patients' survival experiences. Lifetime and age-conditional risks of disease convert incidence rates and mortality from other causes into risk estimates applicable to individuals. Prevalence estimates show the population burden of disease. Over the last twenty years, new measures have been refined to estimate both types of statistics, and methods have been developed to estimate and/or characterize them.

Communicating cancer statistics is important for surveillance research. Various approaches are required to disseminate statistics to different audiences, such as the general public, reporters, policy makers, and researchers. Additionally, different considerations need to be taken into account when reporting large tables of cancer statistics or individual analyses. In particular, summary tables across all ages tend to hide age-specific trends. A new SEER portal with visual presentations that could increase understanding and absorption for different audiences is currently being developed.

Additional developments in statistical methodology in cancer surveillance are also under way. For example, NCI recently developed a new method called CI*Rank to construct confidence intervals for ranking age-adjusted rates across geographic units (usually states or counties). Ranking of health indices provides useful information, but ranks are often viewed as fixed, which may be misleading because the ranking is based on random data. The novel CI*Rank method uses Monte Carlo simulation to find the individual and simultaneous confidence intervals of ranks for health indices data. This software will be available on the NCI website soon.

This review does not include all statistical methods used in cancer surveillance, and other methods are available. For example, life table methods are used to compute the lifetime and age-conditional probability of developing and dying of cancer. The associated software is implemented in DEVCAN (http://surveillance.cancer.gov/devcan/; see Figure 2). The Multiple Primary-Standardized Incidence Ratio (MP-SIR; http://seer.cancer.gov/seerstat/mp-sir.html) methodology is used to perform multiple primary analyses and to test hypotheses that explore theoretical links in the etiologies of two cancers. A defined cohort of persons previously diagnosed with cancer is followed through time to compare their subsequent cancer experience to the number of cancers that would be expected based on incidence rates for the general population. These calculations are available in SEER*Stat. Simulation policy modeling developed by the Cancer Intervention and Surveillance Modeling Network (CISNET; http://cisnet.cancer.gov/) can be used to project future trends and aid in the development of optimal cancer control strategies.

Cancer surveillance data and statistical inferences play important roles in cancer research. The statistical methods used to analyze cancer data are somewhat different from the general mainstream statistical methodology. Often, they are tailored to fit the specific aims of the cancer surveillance framework. In this review, we provide an overview of these data and methods, available software, and references. There is still great need for development of new methods, as data collected from surveillance are expanding and statistical methods are evolving around the needs to analyze and interpret different types of data.

Received 5 February 2013

REFERENCES

[1] Wingo, P. A., et al., *A national framework for cancer surveillance in the United States*. Cancer Causes Control, 2005. **16**(2): p. 151–170.

[2] Swan, J., et al., *Cancer surveillance in the U.S.: can we have a national system?* Cancer, 1998. **83**(7): p. 1282–1291.

[3] Kim, H. J., et al., *Permutation tests for joinpoint regression with applications to cancer rates*. Stat Med, 2000. **19**(3): p. 335–351.

[4] Lerman, P. M., *Fitting segmented regression models by grid search*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 1980. **29**(1): p. 77–84.

[5] Hudson, D., *Fitting segmented curves whose join points have to be estimated*. Journal of the American Statistical Association 1966. **61**: p. 1097–1129. MR0210243

[6] Yu, B., et al., *Estimating joinpoints in continuous time scale for multiple change-point models*. Computational Statistics & Data Analysis, 2007. **51**(5): p. 2420–2427. MR2339003

[7] Fay, M. P., Kim, H. J., and Hachey, M., *On using truncated sequential probability ratio test boundaries for monte carlo implementation of hypothesis tests*. J. Comput. Graph. Stat., 2007. **16**(4): p. 946–967. MR2412490

[8] Feder, P. I., *On asymptotic distribution theory in segmented regression problems–identified case*. The Annals of Statistics, 1975. **3**(1): p. 49–83. MR0378267

[9] Liu, J., Wu, S. Y. and Zidek, J. V., *On segmented multivariate regression*. Stat Sin, 1997. **7**(2): p. 497–525. MR1466692

[10] Kim, J. and Kim, H. J., *Asymptotic results in segmented multiple regression*. Journal of Multivariate Analysis, 2008. **99**(9): p. 2016–2038. MR2466549

[11] Kim, H.-J., Yu, B., and Feuer, E. J., *Inference in segmented line regression: a simulation study*. Journal of Statistical Computation and Simulation, 2008. **78**(11): p. 1087–1103. MR2518299

[12] Clegg, L. X., et al., *Estimating average annual per cent change in trend analysis*. Stat Med, 2009. **28**(29): p. 3670–3682. MR2751730

[13] Ahmedin Jemal, E. P. S., Dorell, C., Noone, A.-M., Markowitz, L. E., Kohler, B., Eheman, C., Saraiya, M., Bandi, P., Saslow, D., Cronin, K. A., Watson, M., Schiffman, M., Henley, S. J., Schymura, M. J., Anderson, R. N., Yankey, D., and Edwards, B. K., *Annual Report to the Nation on the Status of Cancer, 1975–2009, Featuring the Burden and Trends in Human Papillomavirus (HPV)–Associated Cancers and HPV Vaccination Coverage Levels*. JNCI, 2013. **105**(2): p. doi: 10.1093/jnci/djs491.

[14] Kim, H. J., et al., *Comparability of segmented line regression models*. Biometrics, 2004. **60**(4): p. 1005–1014. MR2133553

[15] Schell, M. J., *Identifying key statistical papers from 1985 to 2002 using citation data for applied biostatisticians*. American Statistician, 2010. **64**(4): p. 310–317. MR2758562

[16] Brookmeyer, R. and Damiano, A., *Statistical methods for short-term projections of AIDS incidence*. Stat. Med., 1989. **8**(1): p. 23–34.

[17] Harris, J. E., *Reporting Delays and the incidence of AIDS*. Journal of the American Statistical Association, 1990. **85**(412): p. 915–924.

[18] Pagano, M., et al., *Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data*. Biometrics, 1994. **50**(4): p. 1203–1214.

[19] Kalbfleisch, J. D., Lawless, J. F., and Robinson, J. A., *Methods for the analysis and prediction of warranty claims*. Technometrics 1991. **33**: p. 273–285. MR0970998

[20] Doray, L. G., *UMVUE of the IBNR reserve in a lognormal linear regression model*. Insurance: Mathematics and Economics, 1996. **18**(1): p. 43–57. MR1399864

[21] Midthune, D. N., et al., *Modeling Reporting Delays and Reporting Corrections in Cancer Registry Data*. Journal of the American Statistical Association, 2005. **100**(469): p. 61–70. MR2166070

[22] Clegg, L. X., et al., *Impact of reporting delay and reporting error on cancer incidence rates and trends*. J Natl Cancer Inst, 2002. **94**(20): p. 1537–1545.

[23] Hankey, B. F., et al., *Cancer surveillance series: interpreting trends in prostate cancer–part I: Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates*. J Natl Cancer Inst, 1999. **91**(12): p. 1017–1024.

[24] Etzioni, R., et al., *Cancer surveillance series: interpreting trends in prostate cancer–part III: Quantifying the link between population prostate-specific antigen testing and recent declines in prostate cancer mortality*. J Natl Cancer Inst, 1999. **91**(12): p. 1033–1039.

[25] Huang, L., et al., *Adjusting for reporting delay in cancer incidence when combining different sets of cancer registries*. Biom J, 2013. **55**(5): p. 755–770. MR3097379

[26] Chu, K. C., et al., *A method for partitioning cancer mortality trends by factors associated with diagnosis: an application to female breast cancer*. J Clin Epidemiol, 1994. **47**(12): p. 1451–1461.

[27] SEER*STAT, *Incidence-Based Mortality*, National Cancer Institute.

[28] Feuer, E. J., Merrill, R. M., and Hankey, B. F., *Cancer surveillance series: interpreting trends in prostate cancer–part II: Cause of death misclassification and the recent rise and fall in prostate cancer mortality*. J Natl Cancer Inst, 1999. **91**(12): p. 1025–1032.

[29] Hur, C., et al., *Trends in Esophageal Adenocarcinoma Incidence and Mortality*. Cancer, 2012.

[30] Howlader, N., N. A., Krapcho, M., Neyman, N., Aminou, R., Altekruse, S.F., Kosary, C. L., Ruhl, J., Tatalovich, Z., Cho, H., Mariotto, A., Eisner, M. P., Lewis, D. R., Chen, H. S., Feuer, E. J., and Cronin, K. A., (eds). *SEER Cancer Statistics Review, 1975–2009 (Vintage 2009 Populations), National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2009_pops09/, based on November 2011 SEER data submission, posted to the SEER web site, April 2012*. 2012.

[31] Hakulinen, T. and Tenkanen, L., *Regression-Analysis of Relative Survival Rates*. Applied Statistics-Journal of the Royal Statistical Society Series C, 1987. **36**(3): p. 309–317.

[32] Ederer, F., Axtell, L. M., and Cutler, S. J., *The relative survival rate: a statistical methodology*. Natl. Cancer Inst. Monogr, 1961. **6**: p. 101–121.

[33] Ederer, F. and Heise, H., *Intructions to IBM 650 programmers in processing survival computations. Methodological note no. 10, end results evaluation section. Technical report*. National Cancer Insitute, Bethesda, MD, 1959, 1959.

[34] Hakulinen, T., *Cancer survival corrected for heterogeneity in patient withdrawal*. Biometrics, 1982. **38**(4): p. 933–942.

[35] Cho, H., Mariotto, N. H. Angela B., and Cronin, K. A., *Estimating relative survival for cancer patients from the SEER Program using expected rates based on Ederer I versus Ederer II method*. 2011: Surveillance Research Program, NCI, Technical Report #2011-01, 2011.

[36] Perme, M. P., Stare, J., and Esteve, J., *On estimation in relative survival*. Biometrics, 2012. **68**(1): p. 113–120. MR2909859

[37] Greenwood, M., *The Errors of Sampling of the Survivorship Table*, in *Reports on Public Health and Medical Subjects*, H.M.s.S. Office, Editor 1926: London.

[38] Dickman, P. W., et al., *Regression models for relative survival*. Statistics in medicine, 2004. **23**(1): p. 51–64.

[39] Esteve, J., et al., *Relative survival and the estimation of net survival: elements for further discussion*. Statistics in medicine, 1990. **9**(5): p. 529–538.

[40] Nelson, C. P., et al., *Flexible parametric models for relative survival, with application in coronary heart disease*. Statistics in medicine, 2007. **26**(30): p. 5486–5498. MR2416842

[41] Royston, P. and Parmar, M. K. B., *Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects*. Statistics in medicine, 2002. **21**(15): p. 2175–2197.

[42] Giorgi, R., Payan, J., and Gouvernet, J., *RSURV: a function to perform relative survival analysis with S-PLUS or R*. Computer Methods and Programs in Biomedicine, 2005. **78**(2): p. 175–178.

[43] Pohar, M. and Stare, J., *Making relative survival analysis relatively easy*. Computers in biology and medicine, 2007. **37**(12): p. 1741–1749.

[44] Pohar, M. and Stare, J., *Relative survival analysis in R*. Computer Methods and Programs in Biomedicine, 2006. **81**(3): p. 272–278.

[45] Brenner, H., *Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis*. Lancet, 2002. **360**(9340): p. 1131–1135.

[46] Mariotto, A. B., et al., *Estimates of long-term survival for newly diagnosed cancer patients: a projection approach*. Cancer, 2006. **106**(9): p. 2039–2050.

[47] Yu, B., et al., *Modelling population-based cancer survival trends using join point models for grouped survival data*. J. R. Stat. Soc. Ser. A. Stat. Soc., 2009. **172**(2): p. 405–425. MR2664957

[48] Feuer, E. J., et al., *The impact of breakthrough clinical trials on survival in population based tumor registries*. J Clin Epidemiol, 1991. **44**(2): p. 141–153.

[49] Ghosh, P., et al., *Semiparametric Bayesian approaches to joinpoint regression for population-based cancer survival data*. Computational Statistics & Data Analysis, 2009. **53**(12): p. 4073–4082. MR2744305

[50] Yu, B., et al., *Cure fraction estimation from the mixture cure models for grouped survival data*. Statistics in medicine, 2004. **23**(11): p. 1733–1747.

[51] Yu, B., et al., *CANSURV: A Windows program for population-based cancer survival analysis*. Comput. Methods Programs Biomed., 2005. **80**(3): p. 195–203.

[52] Andersson, T. M. L., et al., *Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models*. BMC medical research methodology, 2011. **11**: p. 96.

[53] Hinchliffe, S. R., et al., *Should relative survival be used with lung cancer data?* Br J Cancer, 2012. **106**(11): p. 1854–1859.

[54] Howlader, N., et al., *Improved estimates of cancer-specific survival rates from population-based data*. J Natl Cancer Inst, 2010. **102**(20): p. 1584–1598.

[55] Cronin, K. A. and Feuer, E. J., *Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival*. Stat Med, 2000. **19**(13): p. 1729–1740.

[56] Feuer, E. J., et al., *The Cancer Survival Query System: Making survival estimates from the Surveillance, Epidemiology, and End Results program more timely and relevant for recently diagnosed patients*. Cancer, 2012. **118**(22): p. 5652–5662.

[57] Hakulinen, T., et al., *Testing equality of relative survival patterns based on aggregated data*. Biometrics, 1987. **43**(2): p. 313–325.

[58] Mariotto, A. B., et al., *Life tables adjusted for comorbidity more accurately estimate noncancer survival for recently diagnosed cancer patients*. J. Clin. Epidemiol., 2013. **66**(12): p. 1376–1385.

[59] Lee, M., et al., *Predicting the absolute risk of dying from colorectal cancer and from other causes using population-based cancer registry data*. Stat. Med., 2012. **31**(5): p. 489–500. MR2880486

[60] Cheng, S. C., Fine, J. P., and Wei, L. J., *Prediction of cumulative incidence function under the proportional hazards model*. Biometrics, 1998. **54**(1): p. 219–228. MR1626801

[61] Feldman, A. R., et al., *The prevalence of cancer: estimates based on the Connecticut Tumor Registry*. New England Journal of Medicine, 1986. **315**(22): p. 1394–1397.

[62] Capocaccia, R. and De Angelis, R., *Estimating the completeness of prevalence based on cancer registry data*. Stat. Med., 1997. **16**(4): p. 425–440.

[63] Merrill, R. M., et al., *Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program*. Int. J. Epidemiol., 2000. **29**(2): p. 197–207.

[64] Mariotto, A., et al., *Complete and Limited Duration Cancer Prevalence Estimates*, L. A. G. Ries, et al., Editors. 2002, National Cancer Institute: Bethesda, Maryland.

[65] Capocaccia, R., *Relationships between incidence and mortality in nonreversible diseases*. Statistics in Medicine, 1993. **12**(24): p. 2395–2415.

[66] Capocaccia, R., *Relationships between incidence and mortality in non-reversible diseases*. Stat Med, 1993. **12**(24): p. 2395–2415.

[67] Gigli, A., et al., *Estimating the variance of cancer prevalence from population-based registries*. Statistical methods in medical research, 2006. **15**(3): p. 235–253. MR2227447

[68] Simonetti, A., et al., *Estimating complete prevalence of cancers diagnosed in childhood*. Stat. Med., 2008. **27**(7): p. 990–1007. MR2420167

[69] Mariotto, A. B., et al., *Long-term survivors of childhood cancers in the United States*. Cancer Epidemiol. Biomarkers Prev., 2009. **18**(4): p. 1033–1040.

[70] Parry, C., et al., *Cancer survivors: a booming population*. Cancer Epidemiol. Biomarkers Prev, 2011. **20**(10): p. 1996–2005.

[71] Mariotto, A. B., et al., *Projections of the cost of cancer care in the United States: 2010–2020*. J. Natl. Cancer Inst., 2011. **103**(2): p. 117–128.

[72] Verdecchia, A., et al., *A method for the estimation of chronic disease morbidity and trends from mortality data*. Stat. Med., 1989. **8**(2): p. 201–216.

[73] Mariotto, A. B. and De Angelis, R., *The Method to Estimate Breast Cancer Prevalence at State Level*, 2008.

[74] Verdecchia, A., De Angelis, G., and Capocaccia, R., *Estimation and projections of cancer prevalence from cancer registry data*. Stat. Med., 2002. **21**(22): p. 3511–3526.

[75] Mariotto, A. B., et al., *Projecting the number of patients with colorectal carcinoma by phases of care in the US: 2000–2020*. Cancer Causes Control, 2006. **17**(10): p. 1215–1226.

[76] Yabroff, K. R., et al., *Economic burden of cancer in the United States: estimates, projections, and future research*. Cancer Epidemiol. Biomarkers Prev, 2011. **20**(10): p. 2006–2014.

[77] *Cancer Facts & Figures* 1953, American Cancer Society: New York.

[78] *Cancer Facts & Figures* 1961, American Cancer Society: New York.

[79] *Cancer Facts & Figures* 2012, American Cancer Society: Atlanta.

[80] Tiwari, R. C., et al., *A new method of predicting US and state-level cancer mortality counts for the current calendar year*. CA Cancer J. Clin., 2004. **54**(1): p. 30–40.

[81] Pickle, L. W., 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**(1): p. 30–42.

[82] Chen, H. S., 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**(4): p. 1091–1099.

[83] Zhu, L., 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**(4): p. 1100–1109.

[84] North, *CINA Deluxe Analytic File*, 2011, North American Association of Central Cancer Registries.

[85] Health Resources and Services Administration, *Area Resource File (ARF), National County-level Health Resource Information Database*. 2011 August 25, 2011; Available from: http://www.arf.hrsa.gov/.

[86] Centers for Disease Control and Prevention, *Behavioral Risk Factor Surveillance System*. [cited 2011 August 25]; Available from: http://www.cdc.gov/brfss/technical_infodata/index.htm.

[87] Centers for Disease Control and Prevention, *Mortality Data*. [cited 2011 August 25]; Available from: http://www.cdc.gov/nchs/deaths.htm.

[88] SAS Institute, *The GLIMMIX Procedure*, in *SAS/STAT® 9.2 User's Guide, Second Edition* 2009: Cary, NC. p. 2250–2255.

[89] Moller, B., et al., *Prediction of cancer incidence in the Nordic countries: empirical comparison of different approaches*. Statistics in Medicine, 2003. **22**(17): p. 2751–2766.

[90] Holford, T. R., *The estimation of age, period and cohort effects for vital rates*. Biometrics, 1983. **39**(2): p. 311–324. MR0714415

[91] Huang, N. E., et al., *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*. Proceedings of the Royal Society of London Series a-Mathematical Physical and Engineering Sciences, 1998. **454**(1971): p. 903–995. MR1631591

[92] Kulldorff, M., *SaTScan User Guide for version 9.0*. 2010.

[93] Kulldorff, M., *A spatial scan statistic*. Communications in Statistics-Theory and Methods, 1997. **26**(6): p. 1481–1496. MR1456844

[94] Tango, T. and Takahashi, K., *A flexibly shaped spatial scan statistic for detecting clusters*. Int. J. Health Geogr., 2005. **4**: p. 11.

[95] Patil, G. P. and Taillie, C., *Upper level set scan statistic for detecting arbitrarily shaped hotspots*. Environmental and ecological statistics, 2004. **11**(2): p. 183–197. MR2086394

[96] Gangnon, R. E. and Clayton, M. K., *Likelihood-based tests for localized spatial clustering of disease*. Environmetrics, 2004. **15**(8): p. 797–810.

[97] Kulldorff, M., et al., *Breast cancer clusters in the northeast United States: a geographic analysis*. American Journal of Epidemiology, 1997. **146**(2): p. 161–170.

[98] Jemal, A., et al., *Geographic variation in prostate cancer mortality rates among white males in the united states*. Ann. Epidemiol., 2000. **10**(7): p. 470.

[99] Fang, Z., Kulldorff, M., and Gregorio, D. I., *Brain cancer mortality in the United States, 1986 to 1995: a geographic analysis*. Neuro. Oncol., 2004. **6**(3): p. 179–187.

Huann-Sheng Chen
Surveillance Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
USA
E-mail address: Huann-Sheng.Chen@nih.gov

Angela B. Mariotto
Surveillance Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
USA
E-mail address: mariotta@mail.nih.gov

Li Zhu
Surveillance Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
USA
E-mail address: Li.Zhu@nih.gov

Hyune-Ju Kim
Department of Mathematics
Syracuse University
USA
E-mail address: hjkim@syr.edu

Hyunsoon Cho
Surveillance Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
USA

Division of Cancer Registration and Surveillance
National Cancer Center
Korea
E-mail address: hscho@ncc.re.kr

Eric J. Feuer
Surveillance Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
National Institutes of Health
USA
E-mail address: feuerr@mail.nih.gov