Published in final edited form as:

Stat Med. 2015 May 20; 34(11): 1953–1964. doi:10.1002/sim.6464.

Age-Period-Cohort approaches to back-calculation of cancer incidence rate

Cheongeun Oh* and Theodore R. Holford†

*Biostatistics, Department of Population Health, New York University, NY 10016, USA

†Department of Biostatistics, Yale School of Public Health, 60 College Street, New Haven, CT 06520, USA

Abstract

A compartment model for cancer incidence and mortality is developed in which healthy subjects may develop cancer, and subsequently die of cancer or another cause. In order to adequately represent the experience of a defined population, it is also necessary to allow for subjects who are diagnosed at death, as well as subjects who migrate and are subsequently lost to follow-up. Expressions are derived for the number of cancer deaths as a function of the number of incidence cases and vice versa, which allows for the use of mortality statistics to obtain estimates of incidence using survival information. In addition, the model can be used to obtain estimates of cancer prevalence, which is useful for health care planning. The method is illustrated using data on lung cancer among males in Connecticut.

Keywords

age-period-cohort model; incidence; mortality; survival; back-calculation

1. Introduction

Incidence, survival and prevalence of disease are important summaries used to describe the impact of a chronic degenerative disease on a population. Cancer incidence and mortality rates are often used to assess time trends in a defined population, but both are subject to limitations that can affect their interpretation. Incidence rates can only be assessed over time if an efficient tumor registry is in place that is able to identify and appropriately classify the cases as they occur. The operation of tumor registries is costly, so very often only a subset of the population is monitored in this way. In the US, for example, southern states, until recently, have not been well represented in the Surveillance Epidemiology, and End Results (SEER) program. Cause-specific mortality rates, on the other hand, are widely available, but their interpretation can be difficult because not only do they depend on the occurrence of cancer, but on survival, as well. Hence, time trends in cancer mortality can be due to changes in exposure to risk factors that cause the disease, as well as improvements in treatment and patient care that might enhance prognosis. Incidence is a better indication of the effects of exposure to agents that affect disease risk because mortality also involves survival, not only because prognosis may change over time but not all subjects will die from the disease. Epidemiologists often study both incidence and mortality in order to obtain a perspective on

cancer trends [1], but the relationship between the two measures of the effect of disease on a population are seldom quantified or studied. Estimation of cancer incidence from cancer mortality offers a potentially useful tool for studying disease trends because it controls for survival trends and provides a better way to look at change that results from etiologic agents.

In a population that is effectively monitored for cancer incidence and mortality, the two rates are necessarily related through the survival experience of patients diagnosed with the disease. Manton and Stallard [2–4] made use of this by developing a compartment model for cancer mortality rates. By assuming a specified functional form for the survival rates and obtaining maximum likelihood estimates of parameters, they were able to make use of mortality data to infer incidence trends, and forecast cancer trends. While this provides a very useful model, in the sense that it enables one to use mortality data, which are much more widely available, it does rely on fairly stringent assumptions on the functional form for the survival experience, which can be difficult to evaluate. We have made use of a similar compartment model to study the relationship between cancer incidence and mortality in Connecticut, which has the oldest tumor registry in the world. Incident cancer cases are first identified to the registry and then followed, enabling one to not only estimate the incidence and mortality rates, but to also determine the survival distribution which links the two.

We have developed a model that quantifies the relationship between cancer incidence, survival and mortality. While the parameters are used to estimate hazard functions that are commonly used in summarizing various aspects of disease in a population, this formulation provides a way for linking incidence and mortality. Hence, we are able to estimate the functional form for survival over time in a population monitored by a tumor registry in order to validate assumptions that might be used when only mortality information is available. This relationship can provide estimates of the number of incident cases when they are unavailable. In the work presented here, we used data that included both incidence and mortality, so that the observed incidence data can be used to assess the adequacy of the estimates based on the proposed.

2. Method

2.1 Model for individual history of cancer

In this method of analysis we consider a somewhat simplified model for the clinical course of cancer. Individuals are identified at the time of cancer diagnosis, which is recorded in the Tumor Registry, and the subsequent clinical course of the disease is followed, resulting finally in death from cancer for some patients. When using data from a cancer registry, registry-based cancer mortality statistics may differ somewhat from those reported by vital statistics that are routinely recorded for a population. However, for most cancers reported by the Connecticut Tumor Registry the difference is small. The purpose of this model is to describe the relationship between these two sources of population based data on the cancer experience of a population. In this section we shall refer to the disease of interest as cancer, but in practice this would be much more precisely defined in terms of cancer diagnosed at a specific site as coded by a standard procedure, such as the International Classification of Disease for Oncology (ICD-O).

2.2 Compartment model for cancer

We shall consider the overall course for cancer to be represented by the compartment model shown in Figure 1. Compartments outlined with heavy lines represent the primary states of the system and the arrows represent the clinical course for the vast majority of cancer patients. All individuals begin life in the healthy state, H . Those who live long enough in the geographic region under consideration may be diagnosed with cancer, S , and those that die of the disease enter the state D_1 . The obvious clinical course of disease for patients who are not cured is from healthy (H) to cancer (S) to cancer death (D_1). However, some patients may die of a cause other than cancer (D_2), either from the healthy or the cancer state. A defined geographic region, such as a country or a state, is never a completely closed system with respect to a human population, in that individuals may continually migrate in and out of the region. Hence, we must also keep in mind other states in a comprehensive model for cancer, such as the one represented in Figure 1 by the in migration compartments outlined with light lines. While the number of individuals that undergo these transitions might represent a small minority of the overall population, it is nevertheless quite possible that these “leaks” in the system could turn out to be quite significant for some cancer sites. We also include out-migration as a type of failure that needs to be included in a model for the progress of cancer, and this absorbing state is represented by D_3 . It is not our intention here to develop an exhaustive demographic model, but rather to focus on those aspects that are likely to account for the vast majority of affected individuals involved in the disease process. Hence, one might regard the type of analysis under consideration here as one which is conditional on the other dynamic factors that determine the age structure in a population over time.

We are primarily interested in describing incidence and mortality trends over time, which can be determined relative to an individual’s birth date or cohort, age, and the date at which, the event occurred or period. First we model the hazard associated with a transition from H (healthy) to S (cancer) in year p for an individual a years old and thus born in year $c=p-a$ and denote it as $\lambda(a, p, c)$. However, we shall initially suppress period and cohort terms for simplicity, and reintroduce them later. Let τ represent age of cancer diagnosis for a subject, and let $\lambda(\tau)$ be the corresponding cancer incidence hazard.

The hazard corresponding to transition to death from a specific cause or absorbing state, D_d ($d=1, 2, 3$) at age t for an individual who entered S at τ is $\mu_d(\tau, t)$. The overall death rate (including out migration) that removes cancer cases from the population is

$$\mu_+(\tau, t) = \sum_d \mu_d(\tau, t)$$

yielding the overall survival curve, or probability that an individual remains alive in the population at age t given diagnosis at τ

$$P(\tau, t) = \exp \left\{ - \int_{\tau}^t \mu_+(u, t) \cdot du \right\}$$

Let $Q_d(\tau, t)$ be the crude probability that an individual will die of cause D_d before age t given a diagnosis at τ , which Chiang [5] showed to be

$$Q_d(\tau, t) = \int_{\tau}^t \mu_d(\tau, \nu) \exp \left\{ - \int_{\tau}^{\nu} \mu_+(u, \nu) \cdot du \right\} \cdot d\nu$$

with simplification arising only in some special cases, such as the case in which Chiang's proportionality assumption holds, i.e., the cause-specific hazards are proportional during specified intervals of follow-up, i.e. $\mu_d(\tau, t) = c_{id} \mu_+(\tau, t)$ or the I^{th} interval, $x_i - \tau < x_{i+1}$, yielding

$$Q_d(\tau, t) = c_{id} \cdot [1 - P(\tau, t)].$$

The crude probability of death can be used to determine the expected number of deaths from cause D_d during the interval (τ, t) . These hazard functions are also the fundamental quantities that can be used to express the net probabilities of failure if a particular cause is eliminated, as well as the relative survival rate.

2.3 Method of inference

Cases are categorized into groups, where the I^{th} group depends on age, period when the case was diagnosed, thus identifying birth cohort. Each of the n_i cases in the group is then followed to determine their survival experience. The survival experience for a particular individual can be summarized by $(t_{ij}, \delta_{1ij}, \delta_{2ij}, \delta_{3ij})$ where t_{ij} is the length of follow-up for the j^{th} individual in group i and δ_{dij} is an indicator for the cause of failure, $d=1, 2, 3$. We can represent the overall likelihood function for the model as

$$l(n_i) \cdot l(t_{ij}, \delta_{1ij}, \delta_{2ij}, \delta_{3ij} | n_i)$$

where $l(n_i)$ represents the likelihood for the incidence component of the model, and $l(t_{ij}, \delta_{1ij}, \delta_{2ij}, \delta_{3ij} | n_i)$ the survival experience for the subjects diagnosed with cancer. We shall consider models in which each likelihood component does not share any of the same parameters, so that we can find maximum likelihood estimates by conducting a separate analysis of incidence and survival.

Under fairly general conditions, Brillinger [6] showed that the distribution of the number of cancer cases observed in a group has a Poisson distribution. Hence, we assume that the number of cancer cases in group i has a Poisson distribution with mean $E[n_i] = \lambda_i T_i$, where $T_i = \sum_j t_{ij}$ represents the total person-years experience for the population giving rise to the cases in group i , usually determined by the midyear population when analyzing vital statistics. The most obvious estimator for the I^{th} incidence rate is the maximum likelihood estimator

$$\hat{\lambda}_i = \frac{n_i}{T_i} \quad (1)$$

In general, however, one can also consider alternative estimators that depend on specific models for time trends such as the age-period-cohort model [7, 8] in which the rate may be expressed as a log-linear model

$$\lambda_i = \exp\{X_i \cdot \theta\}$$

where X_i is a row vector of regressor variables for group i and θ the corresponding parameters to be estimated. Models that simultaneously include age, period and cohort are affected by the well known identifiability problem, however, we are only interested in fitted values which are estimable function of the model parameters, i.e., unique [9]. The estimator in equation (1) can be regarded as a fitted value from a saturated model, while some of the alternative models have the advantage of smoothing estimates and reducing the standard error. Models can also be used to incorporate theories of carcinogenesis [10–15] as well as the distribution of important risk factors for disease in the population [16].

For the analysis of failure times we will obtain estimates of the cause-specific death hazards from fitting a modification of the proportional hazards model in which the failure times have a piecewise exponential distribution with hazard for group i given by

$$\mu_{dij} = \exp\{X_{dij} \cdot \phi_d\} \quad \text{for } t_{[j]} < t \leq t_{[j+1]}$$

where X_{dij} are covariates associated with cause D_d , $t_{[j]}$ is the j^{th} cutpoint for follow-up period, and ϕ_d are the corresponding regression parameters to be estimated. It is well known that the piecewise exponential distribution can provide good approximations for various survival models with reasonable accuracy for a variety of shapes for the hazard.

Holford [17] provides details on methods for obtaining maximum likelihood estimates of model parameters for single causes of failure. These models can be fitted by using standard methods for fitting log-linear models to contingency tables [18–20] which was extended to the case of multiple causes of death by Larson [21]. This can be accomplished by obtaining a set of minimally sufficient statistics for the model parameters for each cause of failure, which in this case is the number of failures from cause D_d in group i during interval j , n_{dij} , and the corresponding sums of the follow-up times, $T_{dij} = \sum [\mathbb{I}(t - t_{[j]} > 0) \cdot \min(t - t_{[j]}, t_{[j+1]} - t_{[j]})]$, where $\mathbb{I}(\cdot)$ is indicator function. The kernel of the log likelihood is equivalent to one formed by assuming that n_{dij} has a Poisson distribution with mean μ_{dij}

$$n_{di} \sim \text{Poisson}(\mu_{dij} = \lambda_{dij} \cdot T_{dij}), \quad (2)$$

where λ_{dij} is the hazard rate. Hence, the necessary parameter estimates can be obtained by fitting models to each cause of failure.

2.4 Robertson and Boyle tabular data for vital statistics

Vital statistics, including data from cancer registries, are often summarized in tables, which are broken down by age and period. In this section, we will consider an alternative approach for representing these data, along with its advantages for the model under consideration. We will assume that the data are broken into equal time intervals by age and period, these typically being either 5- or 10-year categories. We let the index $I (=1, \dots, J)$ represent the age category and $J (=1, \dots, J)$ the period category, and birth cohorts are usually identified by tracing the diagonal cells of a table, $(j-i+I)$. Therefore, in general, an Age-Period-Cohort data set has I rows of age groups, J columns of periods. The $(I+J-I)$ diagonal cells are referred to as cohorts born in the same years. However, adjacent diagonals have years overlapping with the diagonals immediately before and immediately after. Therefore birth cohorts identified in this way are not unique [22]. For example, 40–44 year old cases that occur during 1950–1954 could have been born from 1905–14, and 45–49 year old cases diagnosed during the same years could have been born from 1910–1919.

Robertson and Boyle [23] point out that this overlap of cohort intervals can be eliminated by effectively dividing the resulting squares on the age-period plane in half along the diagonal. Figure 2 traces the follow-up experience for a typical individual over time (broken line), showing that not only the cells along the diagonal are involved in follow-up, but some of the off-diagonal cells are affected, as well. For example, an individual diagnosed in cell $(i-I, j-I)$ could die in the same cell, $(i, j-I)$ or any cell, which is intersected by this broken line. Also shown in Figure 2 is the relationship among the age, period and cohort indices that will be used in our model. Age, period, and cohort are three major variables that characterize temporal factors of lung cancer. However the linear dependency between the indices for age, period and cohort makes the parameters not identifiable.

Note that in each age-period square, there are two cohort triangles represented, one above and one below the diagonal line depending on the year of birth. Each of these cohort bands (indexed by $k^* = 1, \dots, I+J$) has the same width as the age and period intervals, unlike the earlier representation in which the cohort was twice as wide. The lower triangle represents individuals that are born earlier on average than the upper triangle, but in this representation the cohorts are unique, in that individuals who begin in a particular cohort band remain within the band throughout their follow-up. We will estimate the effects of age, period, and cohort as model parameters when fitting (2) using a generalized linear model

$$Y_{ijk^*} = f\left(\frac{n_{ijk^*}}{T_{ijk^*}}\right) = \mu + \alpha_i + \beta_j + \gamma_{k^*} + \varepsilon_{ijk^*}$$

or a log-linear model

$$\log(n_{ijk^*}) = \log(T_{ijk^*}) + \mu + \alpha_i + \beta_j + \gamma_{k^*} + \varepsilon_{ijk^*}$$

where age $i=1, \dots, I$; period $j=1, \dots, J$; and cohort $k^*=1, \dots, I+J$. The fitted values are estimable functions of the model parameters, thus, they are not affected by the identifiability problem. The event frequency is n_{ijk^*} , and T_{ijk^*} is the total population years of exposure in triangle (i, j, k^*) .

2.5 Matrix representation of cohort band

The birth cohort bands shown in Figure 2 identify unique groups of subjects, and their life events, including cancer diagnosis and death, all occurring within the same cohort band. Hence, it is convenient to consider the birth cohort bands individually when formulating a model for cancer incidence and mortality. We say that a birth cohort band is complete if no incident cases can have occurred before the first triangle of the birth cohort band. Obviously, a cohort band starting at age 0 would satisfy this condition, but for many cancers this condition can be approximately met when starting a cohort band at relatively young ages such as 35 or younger. We will assume that the birth cohort bands are complete unless otherwise specified, but in general the cohorts represented by the cells below the main diagonal of an age by period tabulation of rates will be incomplete.

Suppose that there are K triangles in a complete birth cohort band. A subject diagnosed in a particular triangle may die of cancer within the same triangle; or he/she may survive only to die in one of the later triangles. Let c_k represent the number of incident cancer cases in the k th triangle and the vector of incidence cases in the cohort band is $C' = (c_1 \ c_2 \ c_3 \ \dots \ c_K)_{1 \times K}$. Similarly, the vector of cancer deaths, which yield the cohort band's contribution to cancer mortality is $D' = (d_1 \ d_2 \ d_3 \ \dots \ d_K)_{1 \times K}$. These two vectors are related by

$$D = \begin{pmatrix} c_1 Q(1, 1) \\ c_1 Q(2, 1) + c_2 Q(2, 2) \\ c_1 Q(3, 1) + c_2 Q(3, 2) + c_3 Q(3, 3) \\ \vdots \\ c_1 Q(K, 1) + c_2 Q(K, 2) + c_3 Q(K, 3) + \dots + c_K Q(K, K) \end{pmatrix} = Q \cdot C' \quad (3)$$

where $Q(i, j)$ represents the probability that a cancer patient diagnosed in the j th triangle fails in the i th triangle from cause D_i , and Q is the cancer mortality probability matrix defined by the lower triangle matrix

$$Q = \begin{pmatrix} Q(1, 1) & 0 & 0 & \dots & 0 \\ Q(2, 1) & Q(2, 2) & 0 & \dots & 0 \\ Q(3, 1) & Q(3, 2) & Q(3, 3) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ Q(K, 1) & Q(K, 2) & Q(K, 3) & \dots & Q(K, K) \end{pmatrix}_{K \times K}$$

The matrix Q cannot be degenerate since it is a lower triangle matrix and all the diagonal elements, $Q(i, i)$ are nonzero by definition. Hence the inverse Q^{-1} exists, so that

$$C = Q^{-1} \cdot D \quad (4)$$

which corresponds to a method of back-calculating the number of cancer cases from the number of deaths.

In order to derive the elements of the cancer mortality probability matrix we follow the approach described by Ma [23] and assume that the distribution of times for cancer cases occurring within any triangle is uniform, i.e., the joint density for age, and period is $f(a, p) = 2/\Delta^2$ within any triangle, where Δ is the length of age and period intervals. If the intervals are wide then the rates may vary considerably, especially in older age groups, but this approximation can be improved by using a finer grid. In this model we are only interested in describing the aggregate population in a triangle because additional detail on diagnosis and treatment for individuals is generally not available on the death certificate. We will obtain estimates for the cause-specific hazards that only depend on age, period and cohort at diagnosis. The marginal density function for length of follow-up in a triangle, t , is the same for upper and lower triangles, $f(a, p) = 2(\Delta - t)/\Delta^2$. In addition, we assume that the cause-specific hazards only depend on the age-period interval of diagnosis, so that we can repress the age at diagnosis once we have identified the square in question. The probability that a cancer patient who is diagnosed in triangle i of a specified cohort band, dies of cancer in that same triangle is

$$\begin{aligned} Q(i, i) &= \int_0^\Delta Q_1 f(t) dt = \frac{2}{\Delta^2} \int_0^\Delta (\Delta - t) Q_1(t) dt \\ &= \frac{2}{\Delta^2} \left(\Delta \cdot \int_0^\Delta Q_1(t) dt - \int_0^\Delta t \cdot Q_1(t) dt \right) \end{aligned}$$

where $Q_1(t)$ is the crude probability that a cancer patient will die of cancer before t years of follow-up. For subsequent triangles, this expression can be generalized as

$$\begin{aligned} Q(j+1, j) &= \frac{2}{\Delta^2} \cdot \int_0^\Delta t \cdot \left[Q_1 \left(\frac{l+2}{2} \Delta - t \right) - Q_1 \left(\frac{l-2}{2} \Delta + t \right) \right] dt \quad \text{for } l=0, 2, 4, \dots \\ &= \frac{2}{\Delta^2} \cdot \int_0^\Delta t \cdot \left[Q_1 \left(\frac{l-1}{2} \Delta + t \right) - Q_1 \left(\frac{l+1}{2} \Delta - t \right) \right] dt \quad \text{for } l=1, 3, 5, \dots \end{aligned}$$

While incidence rates are more appropriate for understanding disease etiology, prevalence yields the proportion of a population affected by the disease at a particular point in time. For cancer, we have defined a process that does not allow for recovery so that prevalence essentially refers to individuals living with a previous diagnosis of cancer even if they are not experiencing symptoms. Prevalence can be especially useful in health care planning because it gives an indication of the resources that will be required to treat patients. The numerator for prevalence consists of those who have been diagnosed with the disease and are still alive at the specified time, which depends on the survival experience of the cancer

patients. Using an approach similar to that used to determine the probability that a cancer patient dies of the disease, we identify single birth cohort bands, and let b_j denote the number of prevalent cases in the corresponding triangle j , i.e., the total number of cancer patients who are alive at the beginning of triangle j .

Let p_{ij} represent the probability that an incident case diagnosed in triangle j will not fail from any cause before entering triangle i . Using assumptions similar to those for determining the number of cancer deaths, we obtain

$$p_{j+1,j} = \frac{2}{\Delta^2} \int_0^{\Delta} t \cdot \exp \left\{ - \int_0^t \mu_+(u) du \right\} dt$$

where $\Delta_t = \left(\frac{l+1}{2} \right) \Delta - t$ for $l=1, 3, 5, \dots$, and $\Delta_t = \left(\frac{l-2}{2} \right) \Delta + t$ for $l=2, 4, 6, \dots$. In addition, $p_{ii} = 1$, and $p_{ij} = 0$, when $i < j$. The relationship between the number of incidence and prevalent cases can be expressed as

$$B = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_M \end{pmatrix} = \begin{pmatrix} c_1 p_{11} \\ c_1 p_{21} + c_2 p_{22} \\ c_1 p_{31} + c_2 p_{32} + c_3 p_{33} \\ \vdots \\ c_1 p_{K1} + c_2 p_{K2} + c_3 p_{K3} + \dots + c_K p_{KK} \end{pmatrix} \quad (5)$$

Equation (5) can be expressed in matrix form as

$$B = P \cdot C$$

where $P = (p_{ij})_{K \times K}$ is a lower triangle matrix with one's on the diagonal, and the other elements determined by equation (5). Substituting the expression for C derived from the back-calculation method shown in equation (4) yields

$$B = P \cdot Q^{-1} \cdot D$$

Hence, if detailed information is available on the cause-specific survival experience of the population, then population based mortality data can be used to derive estimates of prevalence.

In the previous section we described methods for obtaining maximum likelihood estimates of the cause-specific hazard functions using piecewise exponential models. These can be used to estimate the survival functions and crude probabilities of death from a specific cause, which in turn gives the mortality and prevalence statistics. Hence, our basic strategy is

$$\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3 \Rightarrow \hat{Q}_1(t, \tau) \text{ and } \hat{Q}_1(t, \tau) \Rightarrow \hat{p}_{ij} \text{ and } \hat{q}_{ij} \Rightarrow B \text{ and } D.$$

We treat loss to follow-up or out migration as a special type of competing risk because it can have a large influence on cancer mortality, especially for sites with a relatively good prognosis.

3. Application

To illustrate the method, we apply our procedure to lung cancer data from the Connecticut Tumor Registry (CTR). We used all the male lung cancer cases diagnosed between January 1, 1975 and December 31, 2004 and age [30, 85) at time of diagnosis. These subjects were followed from diagnosis until death, lost to follow-up or truncation in 2004, and deaths due to lung cancer were tabulated, yielding the mortality rates. Length of follow-up was measured in single years. Out of 240,434 cases, a majority, 75.58 % (18,173) eventually died of lung cancer and 24.41% (58,697) died of other causes or were lost during follow-up visits. We grouped age, period and corresponding cohorts by 5-year intervals. Consequently, we have 11 age groups, 6 period groups and 17 non-overlapping cohorts. In parallel, both incidence and mortality cases were summarized in an age-period table divided into equal width five-year intervals. Following non-overlapping cohort bands along each diagonal, we observed the cancer incidence and mortality experience for individual birth cohorts as vectors representing the experience of the series of diagonal triangles that correspond to a particular birth cohort. To validate our method, we used observed incidence cases to compare to number of cases estimated by our model.

We first estimated the hazard functions. Similar to the approach of Manton and Stallard [4], we used a piecewise exponential distribution for fitting survival models. A Kaplan-Meier survival curve plot was utilized to determine appropriate time-cut points to define intervals within which the hazard is constrained to be constant. Because the prognosis of lung cancer is poor, more than 50% of patients died within 1 year and 80% within 5 years, although younger age groups have slightly better survival. A model with four intervals for the piecewise exponential model define by time cut-points (1, 3 and 5 years) was found to have a good fit (not shown here). Little bias was found if we assumed the hazards for lung cancer, μ_1 , and other, μ_2 , causes of death were approximated by a step functions with these cut-points. We indicated intervals [0,1), [1,3), [3, 5), [5, ∞) by a 4-level categorical variable, *CI*. The evaluation of the effects of age, birth cohort, period of death and *CI* was performed by means of a log-linear Poisson model. In order to select a most plausible model, stepwise variable selection was performed. We examined the Freeman-Tukey residuals and tested over-dispersion using a score test [24–25]. The results of fitting a series of models, deviances and degrees of freedom (DF) are presented in Table 1.

CI was the most influential factor followed by the effects of age and cohort for lung cancer. The model Age+Period+Cohort+*CI*+(Age+Period+Cohort)× *CI* stands out as one which is more plausible than most. Although the deviance was still large compared to DF, plotting the Freeman-Tukey residuals did not reveal any systematic departure from the model. There was

no significant excess of over-dispersion for the final model. Similarly, for other causes, we performed stepwise selection and the same model including interactive effects with *CI* stands out. There was an excess of over-dispersion for all models, where the deviances were larger than DF. Using an iterative algorithm for estimating the dispersion parameter [24–25], we fitted an over-dispersed Poisson regression model as the final model. The goodness of fit for this model was adequate, as there was no indication of significant lack of fit. This model was then used when estimating the cancer mortality probabilities matrices, *Q* for both causes of failure. By then using the inverse of *Q*, we estimated the number of incident cases.

In Table 2, the estimated incidence cases are shown using our back calculation approach together with observed cases. We calculated standard errors of the estimated incidences by bootstrapping (resampling with replacement) from the sample data with 10,000 times. While the estimated incident cases were overall in good agreement with the observed data, younger age groups seem to be systematically underestimated whereas older age groups (75–79 and 80–84) tend to be overestimated. This is likely to be a result of incomplete non-overlapping cohort bands and/or misspecification of the model, e.g., no allowance for migration. Figure 3 shows relative differences of observed from estimated incidence cases. The youngest age group (30–34) and earliest period (1975–1979) had fitted values that were mostly lower than the observed values. This is because the cohort bands were started at the first age group and were not quite complete by definition, since our model does not account for cases diagnosed before age 30. Because lung cancer has such poor prognosis, the underestimation at this boundary would be expected to practically disappear rather quickly. This may explain the pattern of greater discrepancies between the observed and estimated incidence counts at the boundaries.

4. Discussion and summary

The compartment model for transition from healthy to cancer to death or failure from a specified cause appears to give a reasonably good fit for the lung cancer data from Connecticut. We reported the data for males, but a comparable analysis for the females gave a similarly good fit to the data. It is perhaps not altogether surprising that this model gives a good fit because this site is so lethal that patients do not have much opportunity to depart from the relatively simple pathways allowed in this model. Further work is needed to explore the adequacy of this model in describing the relationship between incidence and survival in other sites, especially those that have a better prognosis. While the crude probability of death from individual causes can be used to generate expressions for the relationship between incidence and mortality in a comprehensive model for the disease process, these expressions are not necessarily the summaries that are used by epidemiologists to understand the process. Loss to follow-up, *D*₃, is generally not considered to be of primary interest and in most applications these failures are treated as censored observations. Under the assumptions of random censorship ([19], p.40), the survival function for an individual diagnosed at age *t* would become

$$p_{-3}(\tau, t) = \exp \left\{ - \int_t^\tau [\mu_+(\tau, u) - \mu_3(\tau, u)] \cdot du \right\}$$

The net probability of death from cancer is given by [5]

$$q_1(\tau, t) = 1 - \exp \left\{ - \int_t^\tau \mu_1(\tau, u) \cdot du \right\} \quad (7)$$

If causes of death act independently, then the net probability describes a hypothetical population in which causes other than cancer death have been eliminated, and the complement of $q_1(\tau, t)$ would be related to the concept of relative survival. MIAMOD [26, 27], for example, employs relative survival as part of an alternative approach for estimating cancer incidence from mortality data. However, it is generally unreasonable to assume such independence because of the likelihood that a serious disease like cancer would weaken a patient, thus affecting their risk of death from heart and other diseases. The relative survival function [8] was developed to address this issue by relating the survival experience of cancer patients to the experience of the general population. If $\mu_2^*(t)$ represents the mortality hazard at age t for the general population, then the relative survival curve is given by

$$p_{REL}(\tau, t) = \frac{\exp \left\{ \int_t^\tau [\mu_1(\tau, u) + \mu_2^*(u)] \cdot du \right\}}{\exp \left\{ - \int_t^\tau \mu_*(u) \cdot du \right\}}$$

which reduces to the complement of $q_1(\tau, t)$ shown in equation (7) when $\mu_2^*(t) = \mu_1(\tau, t)$. These results indicate that the somewhat simplified process that gives rise to cancer incidence and mortality statistics depends on the basic hazard functions that are used to characterize the survival experience for a particular population. Results are usually not presented in this way, however, which can sometimes obscure the interrelationship among these probabilities. By fitting an overall model to data from a defined population, we can determine the extent to which we have been able to characterize the important features of the process, and not ignored major increments or decrements in the cancer population. While practitioners may find the model parameters to be difficult to interpret in isolation, we see that they can be expressed in terms of the overall, the relative and the cause specific survival curves, which are very familiar and commonly used.

We remark on potential limitations of using data from population-based cancer registries that have often been recognized [29, 30]. Population-based surveys do not lend themselves to estimating cancer incidence because disease occurrence within the survey interval is relatively rare. Therefore, expanding population-based state cancer surveillance data systems hold greater promise for monitoring cancer trends. Given the need to develop a better understanding of incidence trends for diseases like lung cancer and the limited information available to estimate trends for the overall population, we believe that modeling approaches like those we have taken in this study can provide a valuable addition for the population that supplements what is available from cancer registry studies. Back-calculating incidence from mortality rates via reasonably good estimate of survival trends is less prone to error than directly modeling the mortality rate by imposing survival rate independent or constant over time, which is a commonly used approach in other studies [27, 28].

Finally, the hazard functions may be subject to overfitting because of the complexity of the model that includes a large number of parameters relative to the number of samples. A model that is overfit may generally have poor predictive performance. In the proposed method, estimated coefficients from the fitted model are used in the back-calculation equations in calculating a cancer mortality probability matrix (Q). Therefore, overfitting is less of a concern with the proposed method.

We compared our back-calculated estimates with those obtained by the Mezzetti and Robertson [32], which share an interest with our method in using mortality data on a population-based model approach. In brief, Mezzetti and Robertson [31] used a Bayesian hierarchical model to estimate age-specific incidence per 1-year by modeling age-specific mortality while the survival rate was fixed as constant. The proposed method produces more accurate estimates of incidence cases. Mezzetti and Robertson [31] may benefit from modeling 1-year age specific data, which however would be seldom available. In this regard, our method is useful and applicable provided the survival experience is available at an individual-level to project the trend to infer incident trends.

While the approach we have described shows promise as a way to study cancer incidence trends, further work is needed to allow for other factors that might be especially important for sites that have better prognosis. Two potential extensions of our method will be pursued in further work and help to alleviate limitations in readily available vital statistics data. First, misclassification of cause of death and the sensitivity of the results to estimates of survival rates may be of importance. Second, in the current paper, we did not attempt fit mortality or incidence curve which could prove to be useful in filling in data for incomplete cohorts, as well as enabling one to calculate prediction intervals [32].

References

1. Doll R, Muir CSFJ. Introduction Trends in Cancer Incidence and Mortality. Guest Editors. 1994
2. Manton KG, Stallard E. Maximum likelihood estimation of a stochastic compartment model of cancer latency: Lung cancer mortality among white females in the U S. Computers and Biomedical Research. 1979; 12:313–325. [PubMed: 487773]
3. Manton KG, Stallard E. A population-based model of respiratory cancer incidence, progression, treatment, diagnosis and mortality. Computers in Biomedical Research. 1982a; 15:342–360.
4. Manton KG, Stallard E. The use of mortality time series data to produce hypothetical morbidity distributions and project mortality trends. Demography. 1982b; 19:223–240. [PubMed: 7095218]
5. Chiang, CL. Introduction to Stochastic Processes in Biostatistics. Wiley; New York: 1968.
6. Brillinger DR. The natural variability of vital rates and associated statistics. Biometrics. 1986; 42:693–712. [PubMed: 3814721]
7. Fienberg; Mason. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. San Francisco: Jossey-Bass, Inc; 1978.
8. Holford TR. The estimation of age, period and cohort effects for vital rates. Biometrics. 1983; 39:311–324. [PubMed: 6626659]
9. Holford TR, Zhang Z, McKay LA. Estimating age, period and cohort effects using the multistage model for cancer. Statistics in Medicine. 1994; 13:23–41. [PubMed: 9061838]
10. Moolgavkar SH, Venson DJ. A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. Mathematical Biosciences. 1979; 47:55–77.

11. Moolgavkar SH, Dewanji A, Venzon DJ. A stochastic two-stage model for cancer risk assessment. I. The hazard function and the probability of tumor. *Risk Analysis*. 1988; 8:383–392. [PubMed: 3201016]
12. Moolgavkar SH, Luebeck EG. Two-event model for carcinogenesis: biological, mathematical and statistical considerations. *Risk Analysis*. 1990; 10:323–341.
13. Luebeck EG, Moolgavkar SH. Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences, USA*. 2002; 99:15095–15100.
14. Hazelton WD, Clements MS, Moolgavkar SH. Multistage carcinogenesis and lung cancer mortality in three cohorts. *Cancer Epidemiology, Biomarkers & Prevention*. 2005; 14(5):1171–1181.
15. Holford TR, Zhang Z, McKay LA. A model for the effect of cigarette smoking on lung cancer incidence in Connecticut. *Statistics in Medicine*. 1996; 15:565–568.
16. Holford TR. Life tables with concomitant information. *Biometrics*. 1976; 32:587–597.
17. Holford TR. The analysis of rates and of survivorship using log-linear models. *Biometrics*. 1980; 36:299–305.
18. Aitkin, M.; Anderson, F.; Hinde, J. *Statistical Modelling in GLIM*. Oxford: Clarendon Press; 1989.
19. Aitkin M, Clayton D. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*. 1980; 29:156–163.
20. Larson MG. Covariate analysis of competing-risks data with log-linear models. *Biometrics*. 1984; 40:459–469.
21. Case R. Cohort analysis of mortality rates as an historical or narrative technique. *British journal of preventive and social medicine*. *British Journal Preventive Social Medicine*. 1956; 10:159–171.
22. Robertson C, Boyle P. Age, period and cohort models: The use of individual records. *Statistics in Medicine*. 1986; 5:527–538.
23. Ma, Z. PhD Dissertation. Yale University; New Haven, CT: 1995. A model describing the relationships among cancer morbidity, survival and mortality.
24. Dean C. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*. 1992; 87(418):451–457.
25. Breslow NE. Extra-Poisson variation in log-linear models. *Applied Statistics*. 1984; 33:38–44.
26. Verdecchia A, Capocaccia R, Egidi V, Golini A. A method for the estimation of chronic disease morbidity and trends from mortality data. *Statistics in Medicine*. 1989; 8:201–216.
27. Capocaccia R. Relationships between incidence and mortality in non-reversible diseases. *Statistics in Medicine*. 1993; 12:2395–2415.
28. Izquierdo JN, Schoenbach VJ. The potential and limitations of data from population-based state cancer registries. *American Journal Public Health*. 2000; 90(5):695–698.
29. Armstrong BK. The role of the cancer registry in cancer control. *Cancer Causes Control*. 1992; 3(6):569–79.
30. Swan J, Wingo P, Clive R, West D, Miller D, Hutchison C, Sondik EJ, Edwards BK. Cancer surveillance in the U.S.: can we have a national system? *Cancer*. 1998; 83(7):1282–1291.
31. Mezzetti M, Robertson C. A hierarchical Bayesian approach to age-specific back-calculation of cancer incidence rates. *Statistics in Medicine*. 1999; 18(8):919–933.
32. Hakulinen T, Dyba T. Precision of incidence predictions based on Poisson distributed observations. *Statistics in Medicine*. 1994; 13:1513–1523.

Figure 1.
A compartment model for cancer incidence and mortality

Figure 2.

Figure 3.
Relative differences of estimated incident cases from observed incident cases by Ages and Years.

Table 1

Summary of Chi-square analysis from fitting a Poisson model for various age, period and cohort models of lung cancer survival rates for Connecticut males based on individual records

Cause	Model	Deviance	DF	P value
Lung cancer	Age	119343	516	<0.0001
	Period	128987	521	<0.0001
	Cohort	125158	510	<0.0001
	CI	4081	523	<0.0001
	Age+ Period+CI	1462	508	<0.0001
	Period+Cohort+ CI	1533	502	<0.0001
	Age+Cohort+ CI	1313.5	497	<0.0001
	Age+Period+Cohort+ CI	1300.1	492	0.0195
	Age+Period+Cohort+CI+ (Age+Period)× CI	437.7	447	0.6137
	Age+Period+Cohort+CI+(Age+Period+Cohort)× CI	400.1	400	0.4885
Other causes	Age	14070.6	516	<0.0001
	Period	14923.2	521	<0.0001
	Cohort	17639.9	510	<0.0001
	CI	17160.94	523	<0.0001
	Age+ Period+ CI	9299.645	508	<0.0001
	Period+Cohort+ CI	8798.9	502	<0.0001
	Age+Cohort+ CI	9372.295	497	<0.0001
	Age+Period+Cohort+ CI	8368.5	492	<0.0001
	Age+Period+Cohort+CI+ (Age+Period)× CI	456 *	447	0.3712
	Age+Period+Cohort+CI+(Age+Period+Cohort)× CI	397 *	400	0.5273

* extra-Poisson was fitted

Estimated vs. observed incidence cases of lung cancer (estimated values with the bootstrapped standard errors are in the parentheses)

Table 2

Period	1975–79	1980–84	1985–89	1990–94	1995–1999	2000–04
Age						
30–34	10 (6.9; 0.17)	20 (16.9; 1.04)	12 (10.7; 0.96)	16 (12.9; 1.07)	12 (9.7; 0.05)	6 (4.7; 0.01)
35–39	36 (27.4; 0.17)	44 (48.6; 0.20)	54 (60.2; 0.18)	33 (42.6; 0.12)	36 (37.0; 0.11)	24 (26.3; 0.11)
40–44	105 (99.7; 0.49)	90 (105.6; 0.35)	110 (127.7; 0.54)	91 (105.8; 0.44)	105 (124.8; 0.43)	89 (99.1; 0.32)
45–49	266 (257.6; 1.28)	229 (245.9; 0.85)	226 (244.7; 0.94)	206 (221.6; 0.95)	188 (201.0; 0.74)	179 (199.0; 0.76)
50–54	480 (474.4; 2.34)	446 (461.9; 2.03)	406 (434.1; 1.82)	339 (349.8; 1.58)	368 (390.0; 1.48)	352 (375.6; 1.46)
55–59	764 (754.8; 3.57)	850 (889.7; 3.66)	746 (776.0; 3.40)	577 (611.9; 2.52)	568 (606.2; 2.53)	588 (627.8; 2.62)
60–64	1009 (1001.3; 5.08)	1137 (1172.3; 5.53)	1151 (1187.4; 5.4)	944 (982.8; 4.42)	858 (891.4; 3.73)	803 (841.1; 3.58)
65–69	1042 (1035.3; 5.23)	1233 (1270.6; 5.89)	1339 (1373.9; 6.33)	1296 (1328.5; 6.18)	1189 (1222.6; 5.3)	1007 (1037.1; 4.13)
70–74	936 (926.2; 5.10)	1130 (1162.4; 5.76)	1260 (1291.1; 6.36)	1325 (1361.1; 6.68)	1257 (1289.0; 6.29)	1212 (1244.1; 5.83)
75–79	678 (670.2; 3.72)	803 (833.2; 4.46)	935 (968.3; 4.8)	1145 (1177.7; 5.72)	1135 (1174.1; 5.76)	1178 (1212.2; 5.67)
80–84	372 (362.7; 2.29)	422 (433.4; 2.62)	469 (493.0; 3.1)	640 (622.5; 3.89)	744 (741.7; 4.45)	796 (802.3; 4.45)