Two Approaches for Estimating Disease Prevalence from Population-Based Registries of Incidence and Total Mortality

Mitchell H. Gail,[1,*] Larry Kessler,[2] Douglas Midthune,[3] and Steven Scoppa[4]

[1]National Cancer Institute, Division of Cancer Epidemiology and Genetics,
Executive Plaza South, Room 8032, 6120 Executive Boulevard,
MSC 7244, Bethesda, Maryland 20892-7244, U.S.A.
[2]Food and Drug Administration, 1350 Piccard Drive, Rockville, Maryland 20855, U.S.A.
[3]National Cancer Institute, Division of Cancer Prevention,
6130 Executive Boulevard, EPN 344, Bethesda, Maryland 20892-7354, U.S.A.
[4]Information Management Services, 12501 Prosperity Drive, Suite 200,
Silver Spring, Maryland 20904, U.S.A.
*email: mit@cu.nih.gov

SUMMARY. Two approaches are described for estimating the prevalence of a disease that may have developed in a previous restricted age interval among persons of a given age at a particular calendar time. The prevalence for all those who ever developed disease is treated as a special case. The counting method (CM) obtains estimates of prevalence by dividing the estimated number of diseased persons by the total population size, taking loss to follow-up into account. The transition rate method (TRM) uses estimates of transition rates and competing risk calculations to estimate prevalence. Variance calculations are described for CM and TRM as well as for a variant of CM, called counting method times 10 (CM10), that is designed to yield more precise estimates than CM. We compare these three estimators in terms of precision and in terms of the underlying assumptions required to justify the methods. CM makes fewer assumptions but is typically less precise than TRM or CM10. For common diseases such as breast cancer, CM may be preferred because its precision is excellent even though not as high as for TRM or CM10. For less common diseases, such as brain cancer, however, TRM or CM10 and other methods that make stabilizing assumptions may be preferred to CM.

KEY WORDS: Bias of prevalence estimate; Cancer registry; Chronic disease prevalence; Lexis diagram; Precision of prevalence estimate; Prevalence estimation.

1. Introduction

We describe two approaches to using registry data to estimate age- and time-specific disease prevalence, $\pi(c_1, c_2, a, s)$, namely the probability that an individual who is alive at calendar time s and is in the age range $[a, a+1)$ had disease incident in the age interval $[c_1, c_2)$, with $c_2 \leq a$. The quantity $\pi(0, a, a, s)$ is the age-specific point prevalence of disease at time s, including all persons who ever developed disease since birth.

One approach, the counting method (CM), estimates the number of disease survivors in the population (Feldman et al., 1986). A Lexis diagram is helpful in understanding the CM (Figure 1). Consider subjects who are in the age range $[a, a+1) = [50, 51)$ on s = January 1, 1990, and whose cancers were incident in the age interval $[c_1, c_2) = [44, 46)$. These subjects' cancers must have arisen in the parallelogram-shaped region of calendar time t and age x shown in Figure 1. Subject 1 died at age 48 and does not contribute to prevalence at s = January 1, 1990. Subject 2 did survive and was counted. Subject 3 was alive when lost to follow-up at age 47, but one can estimate the chance that this subject survived to date s. The CM counts all subjects like subject 2 who are known to have survived to s and adds an estimate of the number of survivors among those who, like subject 3, were alive when lost to follow-up before s.

A second approach is to use data from disease registries to estimate the various intensity (hazard or transition rate) functions that determine point prevalence. As described by Keiding (1991), a person at calendar time t in the healthy state H may transit to the chronic disease state (e.g., cancer), I, with intensity $\alpha(t, x)$ that may depend on calendar time t or age x. Alternatively, the individual may die (state D) with intensity $\mu(t, x)$ directly from state H. A person in state I is

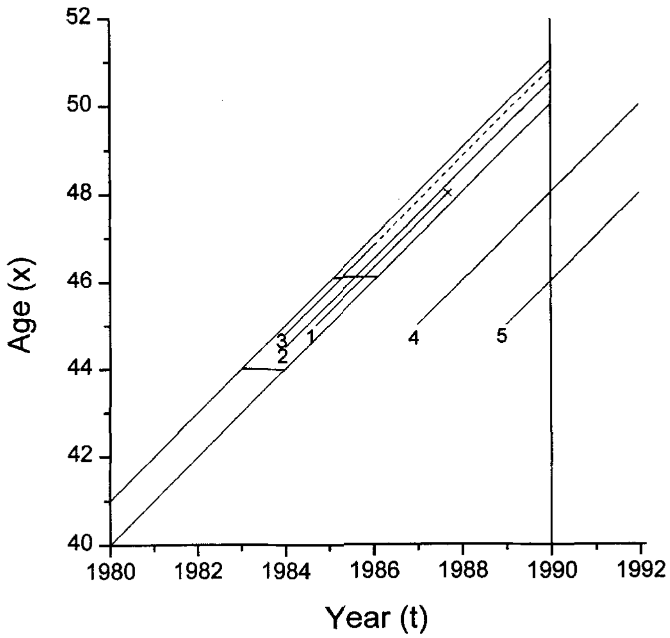

Figure 1. Lexis diagram depicting events in the plane defined by calendar date, t, and age, x. The parallelogram defines the region of cancer incidence for subjects who would be age $[50, 51)$ at calendar time $t = s = 1990$ if they survive and who are ages $[44, 46)$ at cancer incidence. The x symbol indicates the death of subject 1 at age 48. The solid line for subject 3 terminates at age 47, when he is lost to follow-up, and the dotted line indicates his trajectory if he continues to survive to $t = s = 1990$.

at risk of death with intensity $\lambda(t, x, d)$, which may depend on duration d in state I as well as on t and x. These intensities determine the prevalence of the chronic disease if one assumes that the numbers of births at calendar time t is governed by a process with intensity $\beta(t)$ that is independent of the subsequent life histories (Keiding, 1991). We call methods based on modeling the intensities transition rate methods (TRM). In this paper, we make the simplifying assumptions for TRM calculations that $\alpha(t, x) = \alpha(x)$ and $\mu(t, x) = \mu(x)$ depend only on age and that $\lambda(t, x, d) = \lambda(x, d)$ depends only on age and duration with cancer, though these assumptions could be relaxed to allow dependence on calendar time.

The purpose of this paper is to define and compare CM and TRM approaches to estimating prevalence and to present corresponding variance estimates.

2. Methods

2.1 *Prevalence Estimation*

Counting method estimates. Define $n^*(x, t)dtdx$ to be the number of individuals who are at risk of first developing cancer at calendar time $[t, t + dt)$ and age $[x, x + dx)$ and let $N(a, t)$ be the number alive in the population at time t in the age range $[a, a + 1)$. Finally, let

$$S(d; x; t) = \exp\left\{-\int_0^d \lambda(t, x + u, u)du\right\} \qquad (1)$$

be the probability that a person who develops cancer at age x and date t will survive beyond duration d after cancer incidence. Then the desired quantity

$$\pi(c_1, c_2, a, s) = N(a, s)^{-1} \int_a^{a+1} \int_{c_1}^{c_2} n^*(x, s - v + x)\alpha(s - v, x) \times S(v - x; x; s - v)dxdv. \qquad (2)$$

To define CM, let X_i be the exact age at cancer incidence for the ith member of a cancer registry and let T_i be the exact calendar time of cancer incidence for that member. Let Y_i be the exact time of death and U_i be the exact time of loss to follow-up, which is observed if the patient is alive when last followed. We only get to observe the minimum of Y_i and U_i. Let s be the exact calendar date (e.g., January 1, 1990) when prevalence is to be estimated and let $I(\cdot)$ be an indicator function equaling unity when the argument is true and zero otherwise. Then the counting method estimate of $\pi(c_1, c_2, a, s)$ is

$$\begin{aligned}\hat{\pi}_{\mathrm{CM}}(c_1, c_2, a, s) = \Big[&\Sigma I(c_1 \le X_i < c_2, Y_i \ge s, U_i \ge s, \\ &a \le X_i + s - T_i < a + 1) \\ &+ \Sigma\{I(c_1 \le X_i < c_2, Y_i > U_i, U_i < s, \\ &a \le X_i + s - T_i < a + 1)\} \\ &\times \left\{\hat{S}(s - T_i; X_i; T_i)/\hat{S}(U_i - T_i; X_i; T_i)\right\}\Big]/N(a, s), \qquad (3)\end{aligned}$$

where summations are over all members in the registry and $\hat{S}(d; x; t)$ is an estimate of $S(d; x; t)$. The first summation in equation (3) represents cancer cases known to have survived to age a, such as subject 2 in Figure 1, and the second term corresponds to cancer cases who were lost to follow-up before age a, such as subject 3 in Figure 1.

The random variables X_i, T_i, Y_i, and U_i are continuous. In registry data, their values are often truncated to integers. For example, cancers incident in the age range $[j, j + 1)$ are recorded as age j, where j is an integer. In such cases, we take the exact value to be the midpoint of the range of possible values. In the previous example, we would set $X_i = j + 0.5$.

Estimates $\hat{S}(d; x; t)$ are obtained by actuarial methods (Cutler and Ederer, 1958) based on follow-up studies of patients detected in the U.S. Surveillance, Epidemiology, and End Results (SEER) Registry with cancer incident between 1980 and 1989 and with $c_1 \le X_i < c_2$ and $a \le X_i + s - T_i < a + 1$. Because we were only concerned with ages ≤ 50, for which $S(d; x; t)$ depends little on age, all age groups were combined to produce survival curves $\hat{S}(d; \cdot; t)$ that depended on the date of diagnosis but not on age. The curve $\hat{S}(d; \cdot; t)$ was estimated by grouping all individuals whose dates of diagnosis were in the calendar intervals $[1980, 1981)$, $[1981–1985)$, and $[1985, 1990)$. Population sizes were based on 1990 census data.

For uncommon cancers, the estimate $\hat{\pi}_{\mathrm{CM}}(c_1, c_2, a, s)$ may be based on only a small number of incident cases. For example, if $a_1 = 50$, $s = 1990$, $c_1 = 44$, and $c_2 = 46$, as in Figure 1, only cases incident at ages $[44, 46)$ and born in the interval $[1990 - 51, 1990 - 50) = [1939, 1940)$ are included, as indicated in the parallelogram. The area of this parallelogram is two (in units of years squared). Suppose instead that all cases that arise in the calendar interval [1980, 1990) are included, provided cancer is incident at ages [44, 46). The area of the rectangle bounded by the points (1980, 44), (1990, 44), (1990, 46), (1980, 46) is 20. Thus, approximately 20/2 = 10-fold more cases are counted. By counting cases such as individual 4 (Figure 1) who had cancer incident at age $X_i = 45$ at time $T_i = 1987$, we are assuming that this case behaves as if he were born in [1939, 1940) instead of at $T_i - X_i = 1987 - 45 = 1942$. We assign a fictitious birth date of 1939.5 to each person falling outside the parallelogram or, equivalently, we assign a fictitious date of cancer incidence of $T_i^+ = X_i + (s - a - 0.5)$. The calendar time of death and time lost to follow-up are shifted by the same amount, so that $Y_i^+ = Y_i + (T_i^+ - T_i)$ and $U_i^+ = U_i + (T_i^+ - T_i)$. Then this larger number of cases is analyzed according to equation (3) with T_i^+, Y_i^+, and U_i^+ replacing T_i, Y_i, and U_i, respectively. Because this calculation includes 10 times as many cases as the calculation based only on cases who would have been exactly age $[a, a+1)$ at calendar time s if they survived, we call this method the counting method times 10 (CM10) method and the corresponding estimate is denoted $\hat{\pi}_{\text{CM10}}(c_1, c_2, a, s)$. However many of these cases will be censored. For example, with $a = 50$, $s = 1990$, and registry information through 1992, a person aged 45 at diagnosis at $T_i = 1989$ will be assigned a fictitious date of diagnosis $T_i^+ = 1990 - 50.5 + 45 = 1984.5$, but follow-up will end after $1992 - 1989 = 3$ years (see subject 5 in Figure 1). To estimate $S(d; x; t)$ for use with the CM10, we performed actuarial calculations on all subjects with cancer incident in 1980–1989 who satisfied $c_1 \leq X_i < c_2$.

To estimate the prevalence among all individuals who are less than a years old at calendar time s and whose disease developed in the time interval $[s - L, s)$, we compute the weighted average

$$\begin{aligned}
&\hat{K}_{\text{CM}}(a, L)\\
&\quad = \sum_{i=0}^{a-1} \hat{\pi}_{\text{CM}}(i - L, i, i, s) N(i, s) / N(+, s)\\
&\quad = \Big[\Sigma I(s - L \leq T_i < s, Y_i \geq s, U_i \geq s, X_i + s - T_i < a)\\
&\qquad + \Sigma\{I(s - L \leq T_i < s, Y_i > U_i, U_i < s,\\
&\qquad\qquad X_i + s - T_i < a)\}\\
&\qquad \times \big\{\hat{S}(s - T_i; X_i; T_i) / \hat{S}(U_i - T_i; X_i, T_i)\big\}\Big] / N(+, s),
\end{aligned} \tag{4}$$

where $N(+, s) = \Sigma_{i=0}^{a-1} N(i, s)$ is the size at time s of the population age $< a$.

In the examples (Section 3), we set s = January 1, 1990, and $L = 10$, so that SEER cases incident from January 1, 1980, to December 31, 1989, are used for $\hat{\pi}_{\text{CM10}}$ and for $\hat{K}_{\text{CM}}(50, 10)$.

Transition rate estimates. Suppose the transition intensities defined in Section 1 are independent of date of birth, so that they are described by $\mu(x)$, $\alpha(x)$, and $\lambda(x, d)$. The quantity $S(d; x)$ is obtained from equation (1) by suppressing t. Let $S_c(a) = \exp(-\int_0^a \alpha(x)dx)$, $S_d(a) = \exp(-\int_0^a \mu(x)dx)$, and $S^*(a)$ be the probability of surviving to age a in the general population. Then the probability that a member of this population would develop cancer in $[c_1, c_2)$ and survive to age $a + 0.5$ is

$$\begin{aligned}
&\pi_{\text{TRM}}(c_1, c_2, a, s)\\
&\quad = \big\{S^*(a + 0.5)\big\}^{-1}\\
&\qquad \times \int_{c_1}^{c_2} S_d(x) S_c(x) \alpha(x) S(a + 0.5 - x; x) dx.
\end{aligned} \tag{5}$$

Keiding (1991) gives an equivalent expression for prevalence odds except that we restrict cancer incidence to the age interval $[c_1, c_2)$ (see also Capocaccia and De Angelis [1997]). Equation (5) equals equation (2) if the integrand in equation (2) is linear on the small interval $[a, a+1)$ and if there are no secular trends and no immigration or emigration.

To estimate $S^*(s)$ for the examples in Section 3, we use U.S. national age-, sex-, and race-specific mortality rates for all causes of death based on the calendar period 1980–1989. We estimate the race and gender distribution of the SEER population as the average of the distributions from the 1980 and 1990 censuses. From overall U.S. national mortality rates $\mu_i^*(a)$ for the ith race and gender group, we compute the corresponding survival curve $S_i^*(a)$. Then $S^*(a)$ is a weighted average $S^*(a) = \Sigma_i f_i S_i^*(a)$, where f_i is the proportion of the SEER population in race and gender group i. For breast cancer, i only ranges over race. We compute $S_d(a)$ exactly as for $S^*(a)$; however, instead of using the overall mortality rates $\mu_i^*(a)$, we set $\mu_i(a)$ equal to the mortality rate from all causes of death except the cause of interest.

The estimated prevalence $\hat{\pi}_{\text{TRM}}(c_1, c_2, a, s)$ is obtained from equation (5) using $\hat{\alpha}(x)$, $\hat{S}_c(x)$, and $\hat{\lambda}(x, d)$ and with $S_d(a)$ and $S^*(a)$ assumed known without error. Estimators $\hat{\alpha}(x)$ and $\hat{\lambda}(x; d)$ are described in the Appendix.

To estimate the prevalence among those less than age a using the TRM method, we replace $\hat{\pi}_{\text{CM}}(c_1, c_2, a, s)$ by $\hat{\pi}_{\text{TRM}}(c_1, c_2, a, s)$ in the first line to the right of the equal sign in equation (4). We call the resulting estimator $\hat{K}_{\text{TRM}}(a, L)$.

2.2 *Variance Estimation*

Counting method. If individuals have independent small probabilities of cancer incidence or if the birth process feeding the three-state transition model is Poisson and subsequent times to cancer incidence are independent of the times of birth and of each other, then the number of cancers incident in a region of the (x, t) plane in Figure 1 is Poisson (Haberman, 1978; Brillinger, 1986; Keiding, 1991) and the number of such cases who survive to a subsequent age interval $[a, a + 1)$ is also Poisson. Therefore, if there were no loss to follow-up and if the first summation in equation (3) counted M cases, the variance of $\hat{\pi}_{\text{CM}}(c_1, c_2, a, s)$ would be estimated as $M\{N(a, s)\}^{-2}$. In the presence of loss to follow-up, the variance calculation is complicated because the estimated weighting function, $\hat{S}(s - T_i; X_i; T_i) / \hat{S}(U_i - T_i; X_i; T_i)$, takes different values depending on when individuals are lost to follow-up. We propose a bootstrap procedure to estimate one component of the variance. The estimate $\hat{\pi}_{\text{CM}}$ can be written as $\hat{\pi}_{\text{CM}} = M\hat{\xi}_M / N(a, s)$, where $M\hat{\xi}_M$ is the numerator in the right-hand side of equation (3). Here M is the number of cases incident in the age range $[c_1, c_2)$ who would be age $[a, a + 1)$ at calendar time s if they survive and $\hat{\xi}_M$ is the estimated proportion of such cases who were alive at calendar time s.

As M increases, $\hat{\xi}_M$ converges to ξ, say, and $M\,\mathrm{var}(\hat{\xi}_M)$ tends to the limiting variance σ^2. It follows that

$$\begin{aligned}\mathrm{var}(\hat{\pi}_{\mathrm{CM}}) &= E\,\mathrm{var}(\hat{\pi}_{\mathrm{CM}} \mid M) + \mathrm{var}\left\{E\hat{\pi}_{\mathrm{CM}} \mid M\right\} \\ &\doteq E\left\{M/N(a,s)\right\}^2 \sigma^2/M + \mathrm{var}\left\{M\xi/N(a,s)\right\} \\ &\doteq M\left\{N(a,s)\right\}^{-2}\sigma^2 + \xi^2\left\{N(a,s)\right\}^{-2}M. \qquad (6)\end{aligned}$$

This variance can be estimated by substituting $\hat{\xi}_M$ for ξ and by estimating σ^2 from the following bootstrap procedure. Fix m and, for $b = 1, 2, \ldots, B = 100$, choose a random sample of size m from the original M cases with replacement. From bootstrap sample b, compute $\hat{\xi}_{m_b}$. The sample variance s^2 of the quantities $\hat{\xi}_{m_b}$ estimates σ^2/m. Hence, we can estimate σ^2 in equation (6) from s^2m. In our calculations, we set $m = 500$, but quantitatively similar estimates were obtained for $m = 1000$, confirming that $m = 500$ sufficed.

Now consider the quantity $\hat{K}_{\mathrm{CM}}(a, L)$ given by equation (4). We can write $\hat{K}_{\mathrm{CM}}(a,L) = \{M/N(+,s)\}\,\hat{\xi}_M$, where M cases were incident at ages $< a$ and where $\hat{\xi}_M$ is the proportion of those M cases who survived to some age $a' \leq a$ at calendar time s and whose cancer was incident in the interval $[a' - L, a')$ for those attaining age a' at calendar time s. From calculations like those yielding equation (6), we find that $\mathrm{var}\{\hat{K}_{\mathrm{CM}}(a,L)\}$ can be estimated from equation (6) with the following changes: $N(+,s)$ replaces $N(a,s)$; ξ now refers to the limiting proportion above and σ^2 refers to the variance of the modified proportion $\hat{\xi}_M$. The quantity σ^2 is estimated by a bootstrap with sample sizes m as for $\hat{\pi}_{\mathrm{CM}}$.

Finally, consider $\hat{\pi}_{\mathrm{CM10}}$. To be specific, we treat the case $a = 50$, $s = 1990$ in Table 1. This method includes all cases incident at age $[40, 50)$ in $[1980, 1990)$ as described in Section 2.1. We can write $\hat{\pi}_{\mathrm{CM10}} = M\hat{\xi}_M\{10 \times N(a,s)\}^{-1}$, where now M is the number of cases that arise at ages $[40, 49)$ in $[1980, 1990)$ and $\hat{\xi}_M$ is the estimated proportion of those cases who survive to age $a = 50$. The variance of $\hat{\pi}_{\mathrm{CM10}}$ is estimated from equation (6) with the following changes: $10 \times N(a,s)$ replaces $N(a,s)$, the new definition of ξ applies, and σ^2 is estimated by bootstrap resampling of m cases and computing the newly defined $\hat{\xi}_M$ on each bootstrap replication.

Transition rate method. The estimator $\hat{\pi}_{\mathrm{TRM}}$ based on equation (5) depends on $\hat{S}(d;x)$, $\hat{\alpha}(x)$, and its associated survival function $\hat{S}_c$. Other functions in equation (6) are assumed to be known without error. Assuming that S_c and S have piecewise constant hazards, we apply the delta method (Rao, 1975, pp. 385–389) to obtain the variance of $\hat{\pi}_{\mathrm{TRM}}$ and of weighted averages such as $\hat{K}_{\mathrm{TRM}}$ (see Appendix).

3. Results

We illustrate these methods on two cancers with very different incidence and survival rates. Breast cancer was selected because the survival distribution following breast cancer incidence is relatively favorable and because incidence rates are relatively high. The age-specific breast cancer incidence rates per 10^5 person-years in SEER from 1980 to 1989 were 27 for women age 30–34 and 394 for women age 70–74. Brain cancer was selected because the survival distribution is relatively unfavorable and the incidence rate is low (3.3 per 10^5 person-years in the age range 30–34 and 20.2 per 10^5 person-years in the age range 70–74). Both on grounds of incidence rates and survival rates, we expect the prevalence of brain cancer to be much smaller than that of breast cancer.

Table 1 presents data needed to estimate prevalences and their standard errors by CM or CM10 for breast and brain cancers that developed in the previous 10 years. Prevalences at age 50 and for persons less than 50 years old are considered. For example, the CM estimate of the prevalence of breast cancer at age 50, $1473/120{,}543 = 1222 \times 10^{-5}$, is obtained by dividing the estimated number of survivors by the population size. Note the small numbers of brain cancers incident among subjects age 50. The bootstrap estimate of σ^2 is given so that the variances can be calculated from equation (6).

The CM estimate of the prevalence of breast cancer incident in the previous 10 years for 50-year-old women is 1222 per 10^5 women, or about 1.2% (Table 2). The estimated standard error, 31.2, corresponds to a coefficient of variation of 2.6%. The CM10 and TRM methods yield similar estimates, namely 1194 and 1205, respectively, but the standard errors of the latter two estimates are only 10.7 and 10.2, corresponding to coefficients of variation of only 0.9 and 0.9%, respectively.

The CM and TRM methods yield estimates of prevalence in women age <50 of 179 and 183, respectively, with corresponding standard errors 1.44 and 1.29 and corresponding

Table 1
Data for breast and brain cancer

Cancer and procedure	Size of population $N(a,s)$	Number incident cancers, M	Estimated number surviving[a] to a	Estimate of σ^2
Breast				
50, CM	120,543	1,754	1473	0.103
50, CM10	1,205,430	17,927	14,391	0.288
<50, CM	8,805,357	19,137	15,765	0.156
Brain				
50, CM	237,564	122	41.8	0.371
50, CM10	2,375,637	1,427	533.5	0.446
<50, CM	17,738,126	4,805	2,875	0.259

[a] This quantity is $N(a,s)\hat{\pi}_{\mathrm{CM}}$, $N(a,s)\hat{\pi}_{\mathrm{CM10}}$, or $N(a,s)\hat{K}_{\mathrm{CM}}$, depending on the entry.

Table 2
Estimated prevalence (per 10^5 subjects) of breast and brain cancers incident in the previous 10 years (with estimated standard error)[a]

	Estimation procedure		
Age group[a]	CM	CM10	TRM
Breast			
50	1222 (31.2)	1194 (10.7)	1205 (10.2)
<50	179 (1.44)		183 (1.29)
Brain			
50	17.6 (3.25)	22.5 (1.22)	21.7 (1.03)
<50	16.2 (0.307)		16.4 (0.301)

[a] For the CM method, the age group 50 is 50 years old on January 1, 1990, and the age group <50 is less than 50 years old on that date.

coefficients of variation 0.8 and 0.7%. The variability of CM estimates is reduced by the averaging process in equation (4) compared to estimates for a single year of age.

Estimates of the prevalence of brain cancer incident in the previous 10 years for persons age 50 were 17.6, 22.5, and 21.7, respectively, for the CM, CM10, and TRM methods. These numbers are sixfold smaller than for breast cancer. The respective standard errors were 3.25, 1.22, and 1.03, corresponding to coefficients of variation of 18.5, 5.4, and 4.8%. These coefficients of variation are much higher than for breast cancer, reflecting the smaller number of incident brain tumors. For people age <50, the brain cancer prevalence rates for cases incident within the previous 10 years were estimated as 16.2 and 16.4, respectively, for the CM and TRM methods; the respective standard errors were 0.307 and 0.301, and coefficients of variation were 1.9 and 1.8%. The prevalence estimates are 11-fold smaller than for breast cancer, but the coefficients of variation are more than twice as large.

Estimates of prevalence from CM, CM10, and TRM methods agreed well except perhaps for brain cancer at age 50 (Table 2). In Section 4, we discuss differences in assumptions underlying these methods that could contribute to such discrepancies.

4. Discussion

This paper presents three procedures for estimating prevalence: the CM, CM10, and TRM methods. We also develop methods to estimate the variances of these prevalence estimates. In particular, the variance of $\hat{\pi}_{\mathrm{TRM}}$ is estimated from the delta method (Appendix), while the variances of $\hat{\pi}_{\mathrm{CM}}$ and $\hat{\pi}_{\mathrm{CM10}}$ are obtained from the decomposition in equation (6), one term of which is estimated by a bootstrap procedure.

Standard counting process approaches (Aalen and Johansen, 1978; Anderson et al., 1993, Section IV.4) are not applicable because λ depends on age and duration with disease (see Anderson et al., 1993, pp. 678–681) and because survival information has been grouped into actuarial intervals.

Somnier, Keiding, and Paulson (1991) estimated prevalence from a TRM-type calculation. The specific assumptions used to model and estimate various transition rates differed from the assumptions used in this paper, however, and no variance estimates were given. Wun, Merrill, and Feuer (1998) present recursive life table methods for estimating prevalence. Capocaccia and De Angelis (1997) present TRM calculations similar to equation (5) and allow for cures and for an incomplete registry coverage period. None of these papers discusses variance calculations. Verdecchia et al. (1989) develop TRM-like methods to estimate both the incidence rate α and the prevalence from data on cause- and age-specific mortality. They assume that the overall death rates μ and λ are known as well as the death rate from the specific cause of interest, such as breast cancer, following onset of that disease. This approach can

Table 3
Key assumptions

	CM	CM10	TRM
General	1. Registry covers all incident cases 2. Persons with cancer who were alive when lost to follow-up have the same survival as others who were not lost to follow-up	1. Same as for CM 2. All subjects with cancer incident in the 10-year time frame have the same intensities μ, α, and λ as subjects born at $s - a - 0.5$; thus, there are no birth cohort effects in the corresponding range of birth dates	1. Same as for CM 2. Intensities μ, α, and λ are independent of date of birth for the range of birth dates used to estimate these quantities 3. There is no immigration or emigration
Additional assumptions used in the examples	1. S and λ depend only on time since cancer incidence, d, and date of cancer incidence, t	1. S and λ depend only on time since cancer incidence, d	1. S and λ depend only on time since cancer incidence, d, and age at cancer incidence, x

therefore be used when a cancer incidence registry is not available, but the analysis depends heavily on the model and on the ability to accurately assign causes of death. Alho (1992) discusses the impact of secular growth in the population on relations between incidence, prevalence, and duration.

Variance calculations in this paper indicate that the coefficients of variation are small for breast cancer and, even for a rare tumor like brain cancer, the coefficients of variation are modest. These data suggest that coefficients of variation can be reduced appreciably for rare tumors, like brain cancer, by relying on the CM10 or TRM estimates of prevalence rather than on the CM estimate. For cancers with higher incidence rates, such as breast cancer, the variances of the CM10 and TRM estimates will also be smaller than the variance of the CM estimate, but the absolute difference in coefficients of variation will be small.

In many situations, therefore, the choice among these estimators should depend not so much on precision as on an assessment of possible systematic error. In these examples, the CM makes fewer assumptions than the CM10 and TRM methods and is therefore less subject to systematic bias (Table 3). Apart from requirements for completeness of coverage of incident cancers and representative follow-up, CM makes few assumptions. Indeed, if there is no loss to follow-up, CM is nonparametric. Moreover, even assumptions on S and λ are less critical for CM and CM10 than for TRM because S is only used to compensate for loss to follow-up in the CM and there is relatively little loss to follow-up in these examples (9.2% for breast cancers and 10.8% for brain cancers). The CM10 makes the additional assumption that age at birth does not alter μ, α, and λ within the range of birth dates covered in the 10-year calendar interval. TRM also requires that there be no birth cohort effects over the range of dates of birth used to estimate μ, α, and λ. In addition, TRM assumes no immigration into or emigration from the hypothetical birth cohort born at $s - a - 0.5$. In the examples, we assumed that S and λ depended on the time interval since cancer incidence and on age at cancer incidence for TRM. It is important to model S correctly for the TRM because S plays an influential role in equation (5). These considerations suggest that, if one is interested in the prevalence of a common cancer on a given date, such as January 1, 1990, one can rely on the CM method to give estimates that take secular trends in α and μ into account implicitly and thus avoid bias without imposing a great loss of precision. For rare tumors, however, estimators like $\hat{\pi}_{\text{CM10}}$ and $\hat{\pi}_{\text{TRM}}$ may be preferred to improve precision.

All these methods are subject to increased potential for systematic errors for estimating the prevalence of cancers in persons diagnosed more than 10 years earlier as, e.g., in estimating the lifetime prevalence, $\pi(0, 50, 50, 1990)$, as discussed by Capocaccia and De Angelis (1997).

Acknowledgements

We thank Drs Eric Feuer and Lap-Ming Wun and the reviewers and editor for corrections and suggestions that improved the presentation.

Résumé

Deux approches sont décrites pour estimer la prévalence d'une maladie à un âge donné pour une époque donnée parmi les sujets ayant développé la maladie dans une tranche d'âge délimitée antérieure. La prévalence parmi ceux n'ayant pas développé antérieurement la maladie est traitée comme un cas particulier. La méthode de comptage (CM) fournit une estimation de la prévalence en divisant le nombre de malades par l'effectif de la population en prenant compte les perdus de vue. La méthode des taux de transition (MTT) utilise une estimation des taux de transition et la prise en compte des risques compétitifs pour estimer la prévalence de la maladie. Le calcul de la variance est décrit pour chacune des méthodes CM et MTT et pour une variante, appelée CM 10, de la méthode CM qui conduit à des estimations plus précises. Les trois méthodes sont comparées du point de vue de leur précision et des hypothèses sous-jacentes nécessaires à leur application. La méthode CM exige moins d'hypothèses, mais elle est moins précise que les méthodes MTT et CM 10. Pour des maladies fréquentes comme le cancer du sein, on peut préférer la méthode CM qui conduit à une bonne précision, même si elle est moins élevée que pour les deux autres méthodes. Cependant pour les maladies plus rares, comme les tumeurs cérébrales, les méthodes MTT et CM 10 sont préférables à la méthode CM.

References

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics* **5,** 141–150.

Alho, J. M. (1992). On prevalence, incidence and duration in general stable populations. *Biometrics* **48,** 587–592.

Andersen, P. K., Borgan, O., Gill, K. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics (with discussion). *Biometrics* **42,** 693–734.

Capocaccia, R. and De Angelis, R. (1997). Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine* **16,** 425–440.

Cutler, S. J. and Ederer, F. (1958). Maximum utilization of the life table in analyzing survival. *Journal of Chronic Diseases* **8,** 699–712.

Feldman, A. R., Kessler, L., Myers, M. H., and Naughton, M. D. (1986). The prevalence of cancer. Estimates based on the Connecticut Tumor Registry. *New England Journal of Medicine* **315,** 1394–1397.

Haberman, S. (1978). Probabilistic treatment of the incidence and prevalence of disease. *Social Science and Medicine, Series A* **12,** 159–161.

Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective. *Journal of the Royal Statistical Society, Series A* **154,** 371–412.

Rao, C. R. (1975). *Linear Statistical Inference and Its Applications.* New York: Wiley.

Somnier, E. E., Keiding, N., and Paulson, O. B. (1991). Epidemiology of myasthenia gravis in Denmark: A longitudinal and comprehensive population survey. *Archives of Neurology* **48,** 733–739.

Verdecchia, A., Capocaccia, R., Egidi, V., and Golini, A. (1989). A method for the estimation of chronic disease morbidity and trends from mortality data. *Statistics in Medicine* **8,** 201–216.

Wun, L.-M., Merrill, R. M., and Feuer, E. J. (1998). Estimating lifetime and age-conditional probabilities of developing cancer. *Lifetime Data Analysis* **4,** 169–186.

Received April 1997. *Revised October* 1998.
Accepted November 1998.

Appendix

Calculating TRM Prevalence and Its Variance

To estimate $\alpha(x)$, $S_c(x)$, and $\lambda(x,d)$ for cancers diagnosed up to 10 years before 1990, we used SEER incidence data and data on survival following cancer from the years 1980 to 1989 inclusive. To estimate $\alpha(x)$, we define age intervals $i = 1, 2, \ldots$ corresponding to ages $[0, 5)$, $[5-10), \ldots$. The estimate of $\alpha(x)$ for x in the ith age interval is

$$\hat{\alpha}(i) = \sum_{t=1980}^{1989} C_{it} / \sum_{t=1980}^{1989} N^*(i,t),$$

where C_{it} cases are first diagnosed in the SEER Registry in the age interval i in year t and $N^*(i,t)$ is the corresponding number at risk of a first cancer in the SEER population. $\hat{S}_c(x)$ is then computed by assuming $\hat{\alpha}(x)$ is piecewise constant on the age intervals. We estimate $N^*(i,t)$ iteratively by first estimating $\alpha(i)$ with $N^*(i,t) = N(i,t)$ and then repeating the calculation one time with $N^*(i,t) = \hat{S}_c(5i-2.5)N(i,t)$.

We estimate $S(d;x)$ by assuming the corresponding hazard λ_{ij} is constant on 1-year time intervals following cancer diagnosis. Here λ_{ij} is the hazard of death in year i following diagnosis of cancer for individuals diagnosed at age x in the age interval j. Five-year age intervals (j) were used so that we are in fact estimating a set of survival curves, one for each age group. The hazard λ_{ij} is estimated as $\hat{\lambda}_{ij} = -\ln\{1 - D_{ij}/(R_{ij} - 0.5L_{ij})\}$, where D_{ij} cases die in the ith year following cancer diagnosis among those who were in age interval j at the time of diagnosis and where R_{ij} and L_{ij} are, respectively, the corresponding number at risk at the beginning of interval i and the number who are lost to follow-up in this interval.

When the hazard functions $\mu_i(x)$, $\hat{\alpha}(x)$, and $\lambda(a-x;x)$ are constant on the interval $[g_1, g_2)$, $\hat{\pi}_{\mathrm{TRM}}(g_1, g_2, a, s)$ can be integrated analytically as

$$\begin{aligned}
&\hat{\pi}_{\mathrm{TRM}}(g_1, g_2, a, s) \\
&\quad = \frac{1}{S^*(a)} \int_{g_1}^{g_2} (\Sigma_i f_i S_{di}(x))\, \hat{S}_c(x)\hat{\alpha}(x)\hat{S}(a-x;x)dx \\
&\quad = \frac{1}{S^*(a)} \Sigma_i f_i S_{di}(g_1)\hat{S}_c(g_1)\hat{\alpha}(g_1)\hat{S}(a-g_2;g_1) \\
&\qquad \times \left(\frac{1}{\hat{\lambda}(a-g_2;g_1) - \mu_i(g_1) - \hat{\alpha}(g_1)}\right) \\
&\qquad \times \left(e^{-(g_2-g_1)\{\mu_i(g_1)+\hat{\alpha}(g_1)\}} - e^{-(g_2-g_1)\hat{\lambda}(a-g_2;g_1)}\right).
\end{aligned} \tag{A.1}$$

Assume that $\mu_i(x)$, $\hat{\alpha}(x)$, and $\hat{\lambda}(a-x;x)$ are piecewise constant functions, namely, $\mu_i(x) = \mu_{ij}$ for $a_j \le x < a_{j+1}$, $j = 1, \ldots, n_a$, $\hat{\alpha}(x) = \hat{\alpha}_j$ for $b_j \le x < b_{j+1}$, $j = 1, \ldots, n_b$, and $\hat{\lambda}(a-x;x) = \hat{\lambda}_{kj}$ for $d_j \le x < d_{j+1}$, $e_k \le a - x < e_{k+1}$, $j = 1, \ldots, n_d$, $k = 1, \ldots, n_e$. Then we calculate $\hat{\pi}_{\mathrm{TRM}}(c_1, c_2, a, s)$ by breaking the interval $[c_1, c_2)$ into subintervals $[g_m, g_{m+1})$, $m = 1, \ldots, n_g$ such that $\mu_i(x)$, $\hat{\alpha}(x)$, and $\hat{\lambda}(a-x;x)$ are constant on $[g_m, g_{m+1})$:

$$\hat{\pi}_{\mathrm{TRM}}(c_1, c_2, a, s) = \Sigma_{m=1}^{n_g} \hat{\pi}_{\mathrm{TRM}}(g_m, g_{m+1}, a, s),$$

where $\{g_1, \ldots, g_{n_g+1}\}$ is the ordered set with elements $\{x : c_1 \le x \le c_2$ and $(x = c_1$ or $x = c_2$ or $x \in \{a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}, d_1, \ldots, d_{n_d}\}$ or $a - x \in \{e_1, \ldots, e_{n_e}\})\}$.

Similarly,

$$\begin{aligned}
\hat{K}_{\mathrm{TRM}}(a, L) &= \sum_{i=1}^{a} \hat{\pi}_{\mathrm{TRM}}(i-L, i, i, s)N(i,s)/N(+,s) \\
&= \sum_{i=1}^{a} \frac{N(i,s)}{N(+,s)} \sum_{m=1}^{n_{g_i}} \hat{\pi}_{\mathrm{TRM}}(g_{im}, g_{i(m+1)}, i, s),
\end{aligned}$$

where $\{g_{i1}, \ldots, g_{i(n_{gi}+1)}\}$ is the ordered set with elements $\{x : i - L \le x \le i$ and $(x = i - L$ or $x = i$ or $x \in \{a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}, d_1, \ldots, d_{n_d}\}$ or $a - x \in \{e_1, \ldots, e_{n_e}\})\}$.

The variances of $\hat{\pi}_{\mathrm{TRM}}$ and $\hat{K}_{\mathrm{TRM}}$ are

$$\begin{aligned}
&\operatorname{var}(\hat{\pi}_{\mathrm{TRM}}(c_1, c_2, a, s)) \\
&\quad = \sum_{m=1}^{n_g} \sum_{k=1}^{n_g} \operatorname{cov}(\hat{\pi}_{\mathrm{TRM}}(g_m, g_{m+1}, a, s), \hat{\pi}_{\mathrm{TRM}}(g_k, g_{k+1}, a, s))
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{var}\left(\hat{K}_{\mathrm{TRM}}(a, L)\right) \\
&\quad = \sum_{i=1}^{a} \sum_{j=1}^{a} \frac{N(i,s)N(j,s)}{N(+,s)^2} \\
&\qquad \times \sum_{m=1}^{n_{g_i}} \sum_{k=1}^{n_{g_j}} \operatorname{cov}\left(\hat{\pi}_{\mathrm{TRM}}(g_{im}, g_{i(m+1)}, i, s),\right. \\
&\qquad\qquad \left.\hat{\pi}_{\mathrm{TRM}}(g_{jk}, g_{j(k+1)}, j, s)\right).
\end{aligned}$$

We use the delta method (Rao, 1975, pp. 385–389) to estimate covariances such as those being summed in the above equations.

The hazard estimators $\hat{\alpha}_1, \ldots, \hat{\alpha}_{n_b}, \hat{\lambda}_{11}, \ldots, \hat{\lambda}_{n_e 1}, \ldots, \hat{\lambda}_{1 n_d}, \ldots, \hat{\lambda}_{n_e n_d}$ are assumed independent and have estimated variances $\widehat{\operatorname{var}}(\hat{\alpha}_j) = (\Sigma\, C_{jt})/(\Sigma\, N^*(j,t))^2$ and $\widehat{\operatorname{var}}(\hat{\lambda}_{kj}) = D_{kj}/(R_{kj} - 0.5L_{kj} - D_{kj})^2$. The quantities $\Sigma\, C_{jt}$ and D_{kj} are assumed Poisson and $\operatorname{var}(\hat{\lambda}_{kj})$ is estimated by the delta method. The calculation of $\widehat{\operatorname{var}}(\hat{\alpha}_j)$ ignores the small component of variation that arises from adjusting $N(i,t)$ to remove persons with previous cancer diagnoses.

Also, by the delta method,

$$\begin{aligned}
&\operatorname{cov}(\hat{\pi}_{\mathrm{TRM}}(g_{11}, g_{12}, a_1, s), \hat{\pi}_{\mathrm{TRM}}(g_{21}, g_{22}, a_2, s)) \\
&\quad \approx \sum_{j=1}^{n_b} \left[\frac{\partial}{\partial \hat{\alpha}_j} \hat{\pi}_{\mathrm{TRM}}(g_{11}, g_{12}, a_1, s)\right] \\
&\qquad\qquad \times \left[\frac{\partial}{\partial \hat{\alpha}_j} \hat{\pi}_{\mathrm{TRM}}(g_{21}, g_{22}, a_2, s)\right] \operatorname{var}(\hat{\alpha}_j)
\end{aligned}$$

$$+\sum_{j=1}^{n_d}\sum_{k=1}^{n_e}\left[\frac{\partial}{\partial\hat{\lambda}_{kj}}\hat{\pi}_{\mathrm{TRM}}(g_{11},g_{12},a_1,s)\right] \times\left[\frac{\partial}{\partial\hat{\lambda}_{kj}}\hat{\pi}_{\mathrm{TRM}}(g_{21},g_{22},a_2,s)\right]\mathrm{var}(\hat{\lambda}_{kj}).$$

Let h^* be the index such that $a_{h^*} \le g_1 < g_2 < a_{h^*+1}$. Let i^* be the index such that $b_{i^*} \le g_1 < g_2 < b_{i^*+1}$. Let j^* be the index such that $d_{j^*} \le g_1 < g_2 < d_{j^*+1}$. Let k^* be the index such that $e_{k^*} \le a - g_2 < a - g_1 < e_{k^*+1}$. Then, from equation (A.1), for $j < i^*$,

$$\frac{\partial}{\partial\hat{\alpha}_j}\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s) = -\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s)(b_{j+1}-b_j);$$

for $j = i^*$,

$$\begin{aligned}\frac{\partial}{\partial\hat{\alpha}_j}&\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s)\\ &= -\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s)\left((g_1-b_{i^*})-\frac{1}{\hat{\alpha}_{i^*}}\right)\\ &\quad+\frac{1}{S^*(a)}\Sigma_i f_i S_{di}(g_1)\hat{S}_c(g_1)\hat{\alpha}_{i^*}\hat{S}(a-g_2;g_1)\\ &\quad\times\left(\frac{1}{\hat{\lambda}_{k^*j^*}-\mu_{ih^*}-\hat{\alpha}_{i^*}}\right)\\ &\quad\times\left\{\left(\frac{1}{\hat{\lambda}_{k^*j^*}-\mu_{ih^*}-\hat{\alpha}_{i^*}}\right)\right.\\ &\qquad\times\left(e^{-(g_2-g_1)(\mu_{ih^*}+\hat{\alpha}_{i^*})}-e^{-(g_2-g_1)\hat{\lambda}_{k^*j^*}}\right)\\ &\qquad\left.\times -(g_2-g_1)e^{-(g_2-g_1)(\mu_{ih^*}+\hat{\alpha}_{i^*})}\right\};\end{aligned}$$

for $j > i^*$,

$$\frac{\partial}{\partial\hat{\alpha}_j}\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s) = 0;$$

for $j \neq j^*$,

$$\frac{\partial}{\partial\hat{\lambda}_{kj}}\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s) = 0;$$

for $j = j^*$, $k < k^*$,

$$\frac{\partial}{\partial\hat{\lambda}_{kj}}\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s) = -\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s)(e_{k+1}-e_k);$$

for $j = j^*$, $k = k^*$,

$$\begin{aligned}\frac{\partial}{\partial\hat{\lambda}_{kj}}&\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s)\\ &= -\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s)(a-g_2-e_{k^*})\\ &\quad-\frac{1}{S^*(a)}\Sigma_i f_i S_{di}(g_1)\hat{S}_c(g_1)\hat{\alpha}_{i^*}\hat{S}(a-g_2;g_1)\\ &\quad\times\left(\frac{1}{\hat{\lambda}_{k^*j^*}-\mu_{ih^*}-\hat{\alpha}_{i^*}}\right)\\ &\quad\times\left\{\left(\frac{1}{\hat{\lambda}_{k^*j^*}-\mu_{ih^*}-\hat{\alpha}_{i^*}}\right)\right.\\ &\qquad\times\left(e^{-(g_2-g_1)(\mu_{ih^*}+\hat{\alpha}_{i^*})}-e^{-(g_2-g_1)\hat{\lambda}_{k^*j^*}}\right)\\ &\qquad\left.\times -(g_2-g_1)e^{-(g_2-g_1)\hat{\lambda}_{k^*j^*}}\right\};\end{aligned}$$

and for $j = j^*, k > k^*$, $\frac{\partial}{\partial\hat{\lambda}_{kj}}\hat{\pi}_{\mathrm{TRM}}(g_1,g_2,a,s) = 0$.