 *International Journal of Epidemiology* 2000;**29**:197–207

# Cancer prevalence estimates based on tumour registry data in the Surveillance, Epidemiology, and End Results (SEER) Program

Ray M Merrill,[a] Riccardo Capocaccia,[b] Eric J Feuer[c] and Angela Mariotto[b]

**Background** The Connecticut Tumor Registry (CTR) has collected cancer data for a sufficiently long period of time to capture essentially all prevalent cases of cancer, and to provide unbiased estimates of cancer prevalence. However, prevalence proportions estimated from Connecticut data may not be representative of the total US, particularly for racial/ethnic subgroups. The purpose of this study is to apply the modelling approach developed by Capocaccia and De Angelis to cancer data from the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute to obtain more representative US site-specific cancer prevalence proportion estimates for white and black patients.

**Methods** Incidence and relative survival were modelled and used to obtain estimated completeness indices of SEER prevalence proportions for all cancer sites combined, stomach, cervix uteri, skin melanomas, non-Hodgkin's lymphomas, lung and bronchus, colon/rectum, female breast, and prostate. For validation purposes, modelled completeness indices were computed for Connecticut and compared with empirical completeness indices (the ratio of Connecticut based prevalence proportion estimates using 1973–1993 data to 1940–1993 data). The SEER-based modelled completeness indices were used to adjust SEER prevalence proportion estimates for white and black patients.

**Results** Model validation showed that the adjusted SEER cancer prevalence proportions provided reasonably unbiased prevalence proportion estimates in general, although more complex modelling of the completeness indices is necessary for female cancers of the colon, melanoma, breast, cervix, and all cancers combined. The SEER-based cancer prevalence proportions are incomplete for most cancer sites, more so for women, whites, and at older ages. For all cancers combined, prevalence proportions tended to be higher for whites than blacks. For the site-specific cancers this was true for stomach, prostate, cervix uteri, and lung and bronchus (men only). For colon/rectal cancers the prevalence proportions were higher for blacks through ages 59 (men) and 64 (women), and then for the remaining ages they were higher for whites. Prevalence proportions were lowest for stomach cancer and highest for prostate and female breast cancers. Men experienced higher prevalence proportions than women for skin melanomas, non-Hodgkin's lymphomas, lung and bronchus, and colon/rectal cancers.

**Conclusion** The modelling approach applied to SEER data generally provided reasonable estimates of cancer prevalence. These estimates are useful because they are more representative of cancer prevalence than previously obtained and reported in the US.

**Keywords** Cancer, prevalence, burden, cohort, cross-sectional, life table

**Accepted** 8 September 1999

[a] Department of Health Science, College of Health and Human Performance, Brigham Young University, 213 Richards Building, Provo, UT 84602; Division of Epidemiology, Department of Family and Preventive Medicine, University of Utah College of Medicine, UT 84132, USA.

[b] Laboratory of Epidemiology and Biostatistic, Istituto Superiore di Sanita, Viale Regina Elena, 299, 00161, Roma, Italy.

[c] Applied Research Branch, Cancer Control Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, EPN 313, 9000 Rockville Pike, Bethesda, MD 20892, USA.

Prevalence of disease or health-related conditions are measures of primary interest in public health because they identify a level of burden in the population and on the health care system. Such information is useful in public policy debates when considering allocation of health resources and services. Prevalence represents the proportion of new and pre-existing disease cases or attributes in the population during a specified period of time. This statistic incorporates the combined effects of several factors acting on the population, including disease incidence and survival. The term prevalence used in this study treats an individual diagnosed with the disease as a prevalent case until death. This definition is commonly adopted in prevalence studies[1–5] because problems faced by the survivors can be acute, such as recurrence, but there may also be subtle physical and psychological difficulties as a result of treatment.

Population-based cancer prevalence estimates are only complete if obtained from tumour registry data which have been collected over a sufficiently long period of time to capture all prevalent cases of the disease.[2–3,6] Although incomplete prevalence estimates may be appropriate for determining required treatment in the population for diseases which primarily involve short term care,[7] complete prevalence estimates provide a better assessment of the disease burden for those conditions in which recurrence is common and long term physical and psychological care needed. The Connecticut Tumor Registry (CTR) has information on cancer cases from as early as 1935,[8,9] and is the only registry in the US with sufficient follow-up data to directly estimate cancer prevalence. In the mid 1980s, Feldman *et al.* derived prevalence estimates based on 47 years of incidence data from the CTR.[2] These estimates have more recently been updated using 59 years of incidence data.[5] Nevertheless, the use of CTR data to estimate complete prevalence is limited in that Connecticut incidence and survival data may not be representative of the US population, particularly for racial/ethnic subgroups. Neither of these two CTR-based studies reported prevalence estimates for racial/ethnic subgroups.

Since 1973, the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute began actively collecting and reporting cancer incidence and survival data.[10] The nine standard cancer registries in the SEER Program cover about 10% of the US population. These registries are thought to be reflective of the US cancer experience, and are the primary source of national estimates of cancer incidence and survival. Prevalence proportions based on SEER data are of interest because they better reflect US prevalence among racial subgroups in the population. However, such estimates will be biased for many cancer sites because unobserved cases diagnosed before the start of active data collection by the cancer registries in SEER are not included in the prevalence measure.

In the current study, we determine the degree of completeness of SEER prevalence estimates for both white and black cancer patients using the modelling approach of Capocaccia and De Angelis (1997).[11] This approach is validated using CTR data. The SEER prevalence estimates are then computed and reported, adjusted by the modelled index of completeness. The primary aim of this study is to provide more representative prevalence estimates for the US as well as to provide, for the first time using SEER data, prevalence estimates for blacks. The analysis focuses on ten cancer sites (stomach, cervix uteri, melanomas-skin, non-Hodgkin's lymphomas, lung and bronchus, colon/rectum, female breast, and prostate) which represent a wide range of incidence and survival rates, from low incidence and low survival resulting in low prevalence to high incidence and high survival resulting in high prevalence.

## Materials and Methods

### Tumour registry data and selected cancer sites

Prevalence proportion estimates are based on SEER cancer registry data among patients diagnosed in five states (Connecticut, Iowa, New Mexico, Utah, and Hawaii) and four metropolitan areas (Detroit, Atlanta, San Francisco-Oakland, and Seattle-Puget Sound) between 1 January 1973 and 1 January 1994, and followed through 1993 for vital status. We also estimate prevalence using data from the Connecticut Tumor Registry (CTR) between 1 January 1940 and 1 January 1994. The population at 1 January 1994 was estimated by taking the average of the mid year populations obtained from the Bureau of the Census in 1993 and 1994. Although the CTR began keeping records in 1935, we did not use the first five years of data because of quality issues (i.e. underreporting and misclassification). The SEER registries ensure cancer patient ascertainment and diagnostic information by abstracting hospital records, clinical and nursing home records, records from private pathology laboratories and radiotherapy units. Vital status and cause of death were recorded from death certificates.

A diagnosed case contributes to the pool of prevalent cases until death. Only the first primary for a given cancer type is considered. Cases diagnosed by autopsy and death certificate are not treated as prevalent cases and thus excluded from analysis. Incidence rates were computed by single year of age. Population denominators from the US Bureau of the Census required to compute rates were available in 5-year age groups, with single year ages derived by Beers' 'Ordinary' Formula.[12]

We selected the cancer sites shown in Table 1 because, in addition to providing a good representation of various levels of prevalence, they also represent cancers for which extensive screening and prevention efforts have been made in this country.

### Modelled completeness index

If a number of people diagnosed with cancer before a registry began recording data are still alive at the reference time when the prevalence proportion is estimated, the prevalence measure will be underestimated. In contrast to the SEER registries where an underestimation bias occurs for the majority of cancer sites, the CTR is sufficiently old so that the prevalence proportion includes essentially all diagnosed cases. A method has recently been developed for measuring the underestimation bias of prevalence when computed in relatively young cancer registries, such as those participating in the SEER Program. A description of this general methodology can be referred to elsewhere.[11]

The general methodology considers a single birth cohort with the completeness index a combination of incident and survival functions. Now consider that the birth cohort is observed for a time period of $L$ years. The proportion of the population of individuals with cancer at age $x$ may be separated into a part which derives from the incident cases observed in a registry between the age interval $[x\text{-}L, x]$, and a part of unobserved cases in the registry diagnosed at previous ages and still living at $x$;

**Table 1** Invasive cancer sites considered in this study, according to their incidence and survival rate combinations[a]

| | Survival | | |
|---|---|---|---|
| **Incidence** | Poor (5-year RSR <30%) | Medium (5-year RSR 30–80%) | Good (5-year RSR >80%) |
| Low | Stomach | Cervix | Melanomas-Skin |
| (AIR <20) | | Non-Hodgkin's lymphomas | |
| Medium | Lung & Bronchus | Colorectal | |
| (AIR 20–100) | | | |
| High | | | Breast (female only) |
| (AIR >100) | | | Prostate |

[a] Age-adjusted incidence rates (AIR) and 5-year relative survival rates (RSR), as reported in the SEER Cancer Statistics Review, 1973–1994.[10]

that is, the prevalence at age $x$ consists of the unobserved and observed cancer cases living in the registry:

$$N(x) = N_u(x,L) + N_o(x,L) = \int_0^{x-L} I(t)S(t,x-t)dt + \int_{x-L}^{x} I(t)S(t,x-t)dt$$

where $I(t)$ is the incidence hazard for the disease of interest at the time of diagnosis and $S(t, x\text{-}t)$ is the relative survival[13] from the time of diagnosis to age $x$. A measure of the observed prevalence relative to the total prevalence is called the completeness index, expressed as

$$R = \frac{N_o(x,L)}{N(x)}$$

The completeness index, $R$, varies between zero and one, where one means that all the diagnosed cases were included in the prevalence estimate during the reference period.

## Incidence functions

The incidence function used for describing the relationship between cancer incidence and age, adjusting for birth cohort, is expressed as

$$I(x,k) = \exp(a_k)x^b$$

where $a_k$ is a categorical birth cohort variable coded with k = 16 levels (1888–1892, 1893–1897, ..., 1958–1962, 1963–1968), $x$ is the current age, and $b$ is the slope parameter. The validity of this function has been previously determined to have a biological rationale for a general class of cancers, given the multistage theory of carcinogenesis.[14,15] A linear relationship between log(incidence) and log(age) is obtained by taking the log of both sides of the equation. When the incidence of disease is sufficiently rare, as is true for the cancer sites we consider, this expression can be approximated by a logistic model:

$$I(x,k) = (1 + \exp - (a_k + b\log(x)))^{-1}$$

where $a_k$ is the logit of incidence at the $k$ birth cohort when age equals zero. Hence we were able to use standard logistic regression software which made the estimates easier to compute. This model provides a good fit to the data for each of the cancer sites except all cancers combined, cervix uteri, female breast, and prostate cancers. For these sites the modelled incidence age curve employed a logistic function having as argument a sixth degree polynomial function of age:

$$I(x,k) = \left\{1 + \exp - [a_k + \sum_{i=1}^{6} b_i(x/30 - x_0/30)^i)]\right\}^{-1}$$

where $a_k$ is the logit of incidence at the $k$ birth cohort when age equals the reference age (chosen to be 55), and the constant 30 is an arbitrary scale factor used to avoid very large numbers and numerical instability which arise when taking powers. The predictions used for this method have been previously considered as suitable for many cancer sites, particularly those in which tumour progression and growth are modulated by hormonal factors.[4]

The models were fit to strata defined by various combinations of area (SEER, Connecticut only), gender, and race (white, black) to obtain estimates of the incidence slope parameter $b$ used to derive the completeness index, $R$. Because prevalence for a specific age group is estimated from a single cohort, the numerator and denominator of the prevalence calculation for the completeness index are both scaled by the same cohort parameter, and thus the cohort parameter cancels out of the calculation. The incidence slope estimates were similar between white and black males and white and black females for all cancers combined, stomach, non-Hodgkin's lymphomas, lung and bronchus, and colon/rectum, so the estimated values used to compute the completeness indexes for these sites were based on white and black cases combined. For melanomas, which are very rare among blacks, we only conducted the analysis for whites.

Incidence slope parameter estimates were derived from SEER and Connecticut data by cancer site and gender, for the first incidence function, where $b$ is the slope parameter measuring the log-log linear relationship between incidence and age (estimates not shown). Incidence slope parameter estimates were also derived for the cancer sites assessed using the second incidence function, where $b_i$ is the slope parameter of a polynomial logistic relationship between incidence and age (estimates not shown). The cohort parameters were estimated along with the slope parameters, but as noted they cancel out of the calculation of the completeness index.

## Modelling survival

A survival model with cure was fit to the SEER data. Similar models have been successfully applied previously.[16,17] This model assumes that only a portion of the patients have an excess mortality rate while the remainder have the same mortality

rate as the general population and can be considered with regard to the death risk as cured.

In the survival model, a Weibull function was assumed for fatal cases and the influence of time of diagnosis and race was modelled with an exponential factor of the entire relative survival function. The cumulative relative survival up to age $x$ of a patient diagnosed at age $t$ and year $y$ was assumed as:

$$S(t,x) = [(1 - A) + A\exp(-(\lambda(x - t))^{\gamma})]^{\exp(\beta_1(t-t_0)+\beta_2(y-y_0)+\delta)}$$

where $t_0$ is the reference age and $y_0$ is the reference time. The parameter A represents the proportion of fatal cases (i.e. they are bound to die from the disease) in the patient population, and $\lambda$ and $\gamma$ are the respective scale and shape parameters of the Weibull distribution. The parameters $\beta_1$ and $\beta_2$ are the log relative risk of being diagnosed one year older and one year later, respectively. Finally, $\delta$ is the log relative risk of being black. Raising the survival function to a power of the parameters, as shown, gives a proportional hazards formulation.[18] The parameters A, $\lambda$, and $\gamma$ retain their meaning only for the reference that we fixed at the median age 62 (except for prostate in which we used age 72), the period 1988–1993, and white race. The mean survival time for fatal cases is computed as $1/\lambda\,\Gamma(1/(1+\gamma)$ where $\Gamma$ is the gamma function.

Relative survival estimates using the life table method were computed using the SEER Portable Survival System.[19] Relative survival was stratified according to 5-year age intervals (25–29, 30–34, ..., 75–79, 80–84; except for all cancers, which included earlier age intervals), period of diagnosis (1973–1977, 1978–1982, 1983–1987, 1988–1993), gender, and race (white and black). A 20-year follow-up period was considered and we constrained the relative survival not to increase over follow-up time. Exclusions were made of cases with second or later primaries, cases diagnosed by death certificates and at autopsy, and cases not actively followed. Files containing the empirical relative survival rates and standard errors, together with the corresponding values of time since diagnosis, age, time of diagnosis, gender, and race were exported from the portable survival package. The parameters A, $\lambda$, $\gamma$, $\beta_1$, and $\beta_2$ were then estimated using the SAS NLIN procedure from the exported relative survival results (estimates not shown). On the basis of these parameter estimates, cancer site, gender, and race specific completeness indices were derived.

Figure 1 provides an illustration of 3-, 5-, 7- and 10-year modelled and observed relative survival for white women with breast cancer, presented by year of diagnosis in Connecticut and SEER for the reference age 62. We only use SEER data from 1973 through 1993 to model survival. However, modelled estimates from these years and those back-projected to earlier years (not shown) are used in the denominator of the completeness index R. For Connecticut, where we have observed survival to compare with projected survival from 1940 onward, we compared back-projected and observed survival for breast cancer. The projected survival underestimates the observed survival in early years. This notwithstanding, modelled total prevalence of breast cancer is very close to the observed, as shown in Table 3. This indicates that estimation of completeness indices is scarcely

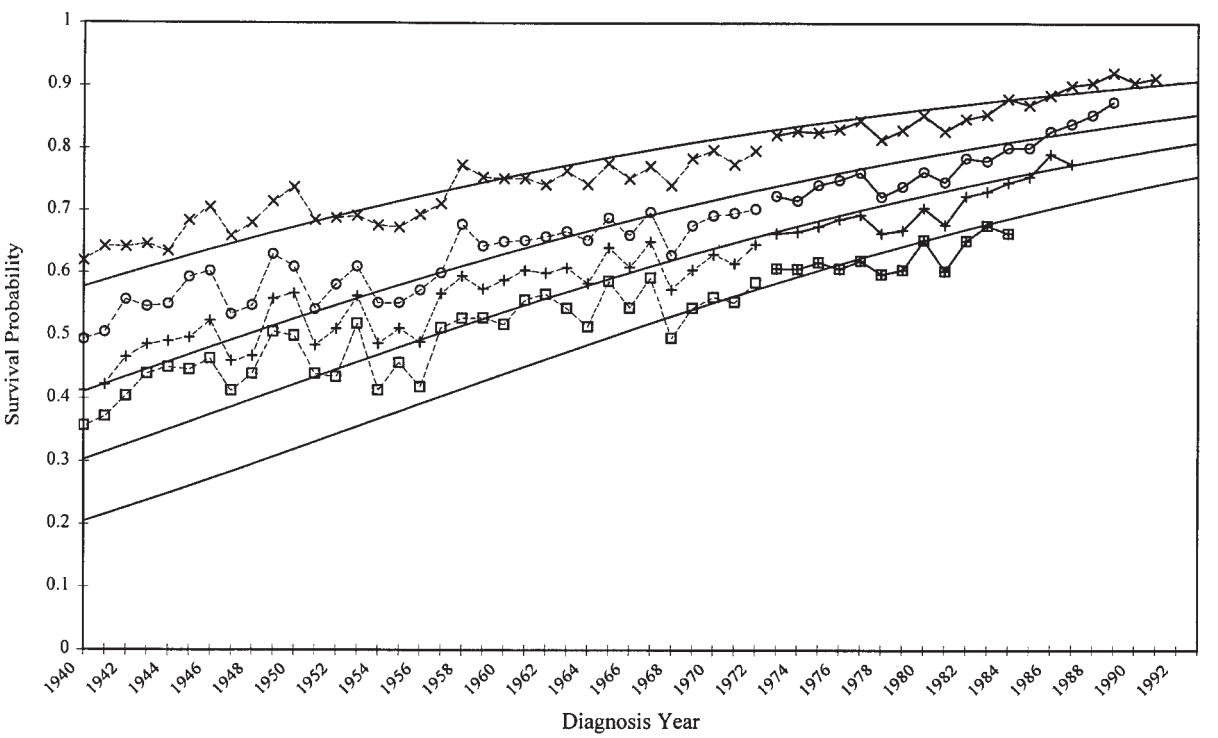


**Figure 1** Breast cancer (invasive only) three, five, seven, and 10 year modelled and observed relative survival for white women in Connecticut at reference age 62,[a] by year of diagnosis[b]

sensitive to the values of the survival function at ⩾20 years before the index date.

### Validation of completeness indices

In order to validate the modelled completeness indices, we considered prevalence estimates computed using the Feldman *et al.* method, which involves using a longstanding registry (i.e. the Connecticut Tumor Registry) to derive the number alive on a certain date, and then making an adjustment for cases lost to follow-up.[2] The life table method was used to estimate the number who would have survived until 1 January 1994 among those lost to follow-up. Life tables were estimated for five cohorts of cases (1940–1953, 1954–1963, 1974–1983, and 1984–1993). The ratio between prevalence based on 1973–1993 data to prevalence based on 1940–1993 was used as an empirical estimate of completeness, and computed for whites by cancer site and gender. We then compared these empirical completeness indices with the corresponding indices obtained through the modelling effort, which we refer to as modelled completeness indices. Modelled completeness indices were obtained for Connecticut whites for validation purposes. Modelled completeness indices were also computed for SEER whites and blacks. We did not attempt to compute modelled completeness indices for Connecticut blacks because of sparse data.

### Adjusted prevalence proportions

Cancer site and age-group specific prevalence proportions on 1 January 1994 were estimated per 100 000 using the Feldman *et al.* method.[2] This method was applied to four data sets: (1) Connecticut white cases diagnosed 1940–1993, (2) Connecticut white cases diagnosed 1973–1993, (3) SEER white cases diagnosed 1973–1993, (4) Connecticut black cases diagnosed 1940–1993, and (5) SEER black cases diagnosed 1973–1993. The method applied to (1) and (4) provides the conventional prevalence proportion estimates. We compare these estimates with those obtained by applying the Feldman *et al.* method to (2), (3), and (5), which were adjusted by dividing the cancer site, age group, area-, and race-specific computed prevalence proportions by the corresponding modelled completeness indices obtained from the Capocaccia and De Angelas method.[11]

## Results

Empirical and modelled completeness indices for all cancer sites combined based on 21 years of follow-up for whites in Connecticut and whites and blacks in SEER (modelled only) are reported by age and gender (Figure 2). The empirical and modelled completeness indices for whites in Connecticut are similar for men but those modelled are lower than the empirical for women. The modelled completeness indices for whites in SEER are similar to the modelled completeness indices in Connecticut. The indices vary according to age and gender, with black men and women in SEER having higher modelled completeness indices than white men and women across all age groups.

Site-specific completeness indices for whites in Connecticut are presented in Figure 3. They tend to be similar or lower (colon/rectal cancers, and melanomas) for women than men, particularly in the older ages. The degree of completeness varies by cancer site and decreases with age. Only for prostate cancer are 21-years of follow-up sufficient to achieve a completeness

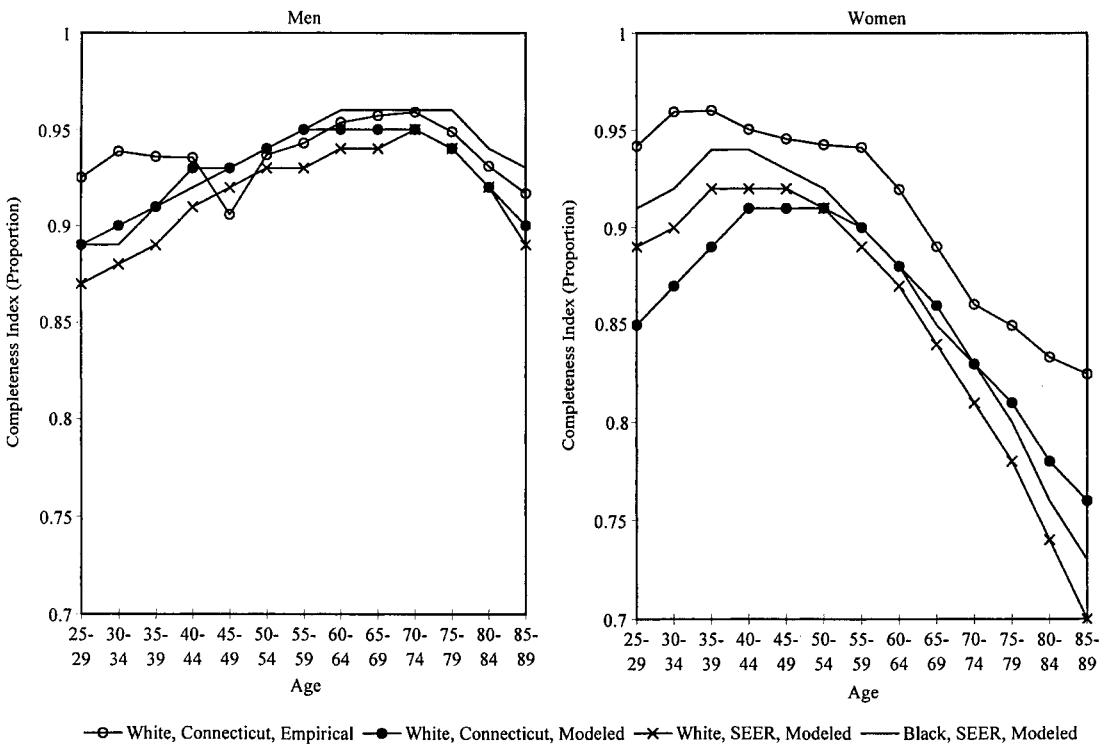


**Figure 2** All cancers combined. Empirical and modelled completeness indices with 21-years of follow-up for whites in Connecticut and whites and blacks in SEER by age and gender

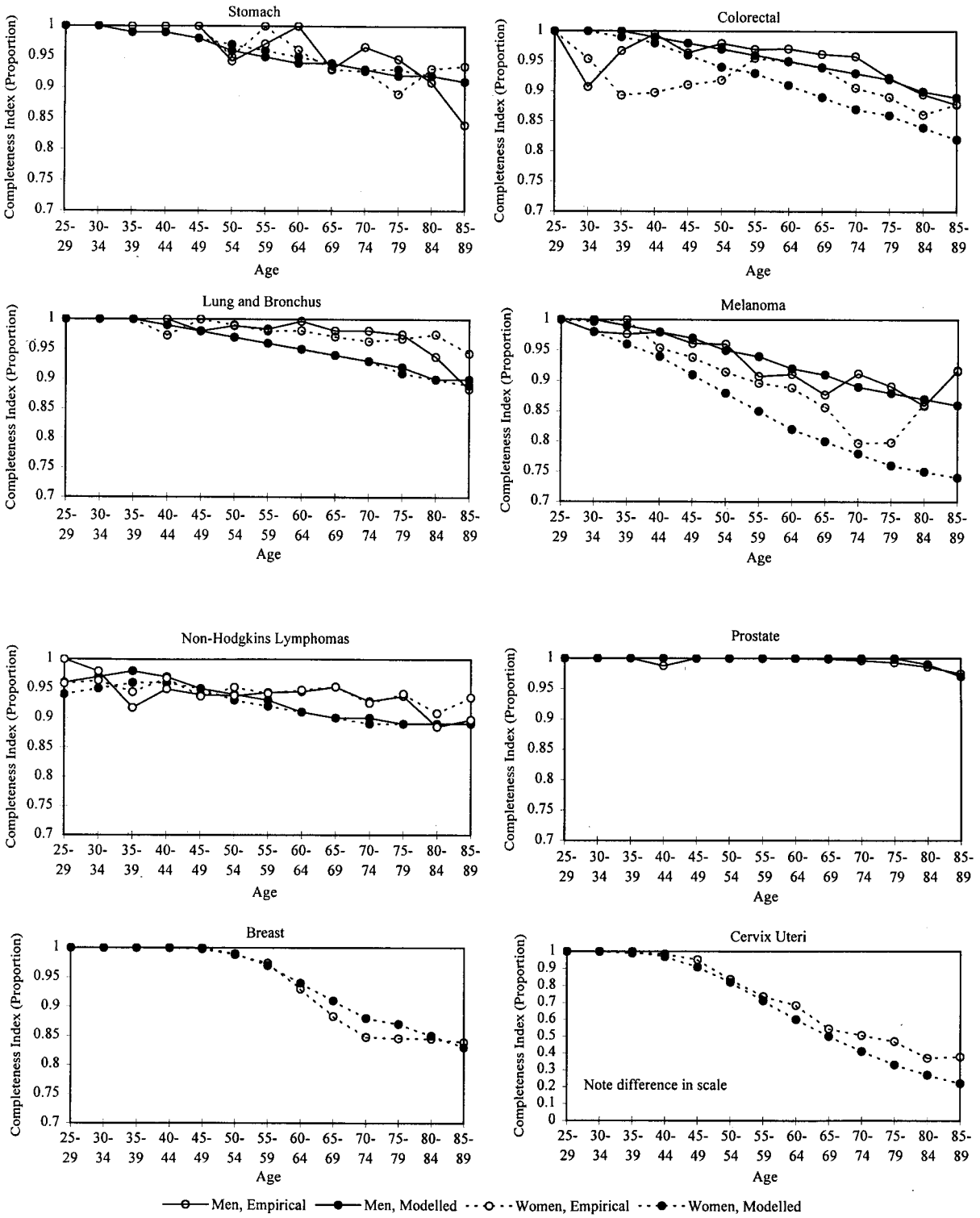


**Figure 3** Empirical and modelled completeness indices with 21-years of follow-up for whites in Connecticut by cancer site, age, and gender

index of unity, except in the last age category 85–89. Cervical cancer presents the lowest level of completeness, falling below 50% for ages over 65. For the other sites, 21 years of follow-up explain a fraction of prevalence ranging between 75% and 95%, even in the oldest age groups. A comparison of the empirical and modelled completeness indices indicates that the modelled indices generally capture the level of completeness for these cancer sites, with a few exceptions. The modelled completeness indices are lower than the empirical completeness indices among women in colon/rectal cancers for ages 30–49, in melanomas for ages 80–89, and in cervical cancer for ages 70–89.

The completeness indices were typically higher for blacks than whites for colon/rectum, lung and bronchus, non-Hodgkin's lymphomas, and female breast cancers (Figure 4). The completeness indices were similar between whites and blacks across age groups for stomach, melanomas, prostate, and cervix uteri cancers (data not shown).

Cancer site and age group specific prevalence proportions on 1 January 1994 per 100 000 are reported for men (Table 2) and women (Table 3). A comparison of prevalence proportion estimates in (1) and (2) gives the relationship between conventional and modelled estimates for whites in Connecticut. This is a measure of the validity of the modelled prevalence estimates. Comparing prevalence proportions in (2) and (3) allows us to identify differences in the modelled prevalence estimates between Connecticut and SEER. This demonstrates the representativeness of Connecticut prevalence to SEER. Comparing prevalence proportion estimates in (1) and (4) or (3) and (5) show differences between estimates for whites and blacks in Connecticut and SEER, respectively.

For men, conventional and modelled prevalence estimates appeared similar for all cancers combined and for each of the selected cancer sites. For women, modelled prevalence estimates tended to be higher than conventional prevalence estimates for all cancers combined, and higher in the older age groups for melanomas, colon/rectum and cervix uteri cancers. Modelled prevalence estimates were generally higher for whites in Connecticut for stomach, non-Hodgkin's lymphomas, and colon/rectal cancers, and higher for whites in SEER for prostate and cervix uteri cancers. Conventional and modelled prevalence estimates were similar between Connecticut and SEER areas for breast, lung and bronchus cancers.

For all cancers combined, prevalence estimates tended to be higher for whites than blacks. For the site-specific cancers this was true for stomach, prostate, cervix uteri, and lung and bronchus (men only). For colon/rectal cancers the prevalence estimates were higher for blacks through ages 59 (men) and 64 (women), and then for the remaining ages they are higher for whites.

As expected, based on the incidence and survival rate combinations reported in Table 1, stomach cancer for men and women had the lowest prevalence estimates whereas prostate and breast cancers had the highest prevalence estimates. Higher prevalence estimates were experienced by men than women for skin melanomas, non-Hodgkin's lymphomas, lung and bronchus, and colon/rectal cancers.

## Discussion

Although the generic survival function used for all of the cancer sites did not fit the data as well as could possibly be done using

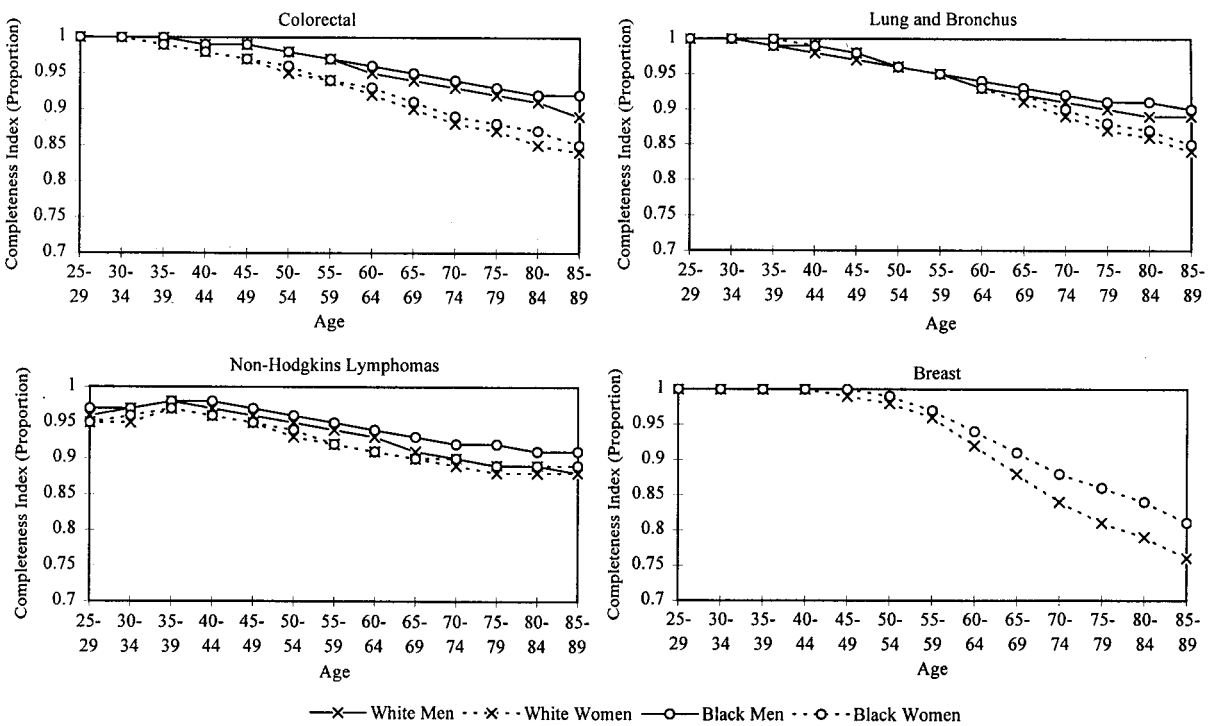


**Figure 4** Modelled completeness indices with 21-years of follow-up for whites and blacks in SEER by cancer site, age, gender, and race

**Table 2** Cancer Site and Age Group Specific Prevalence Proportion Estimates for Men per 100 000 on 1 January 1994[a]

| | Age Group | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cancer Site (Invasive)** | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 | 65–69 | 70–74 | 75–79 | 80–84 | 85–89 |
| **All Cancers** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 436 | 592 | 690 | 978 | 1464 | 2336 | 3809 | 6872 | 10 867 | 14 478 | 18 133 | 21 159 | 23 025 |
| Connecticut Whites, 1973–93[b] | 454 | 617 | 710 | 984 | 1426 | 2328 | 3780 | 6898 | 10 948 | 14 615 | 18 300 | 21 409 | 23 454 |
| SEER Whites, 1973–93[b] | 409 | 603 | 770 | 1019 | 1378 | 2224 | 3807 | 6691 | 11 155 | 15 536 | 19 516 | 22 017 | 23 957 |
| Connecticut Blacks, 1940–93 | 169 | 265 | 380 | 666 | 1183 | 2272 | 3752 | 6781 | 10 846 | 14 511 | 14 656 | 17 870 | 17 827 |
| SEER Blacks, 1973–93[b] | 194 | 274 | 377 | 634 | 1026 | 2134 | 4105 | 6773 | 10 724 | 15 971 | 17 245 | 19 794 | 18 690 |
| **Stomach** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 0 | 2 | 2 | 6 | 11 | 23 | 56 | 89 | 127 | 190 | 208 | 329 | 391 |
| Connecticut Whites, 1973–93[b] | 0 | 2 | 2 | 6 | 11 | 23 | 57 | 94 | 126 | 198 | 214 | 325 | 361 |
| SEER Whites, 1973–93[b] | 1 | 1 | 2 | 4 | 9 | 20 | 40 | 62 | 109 | 160 | 201 | 255 | 280 |
| Connecticut Blacks, 1940–93 | 0 | 0 | 0 | 23 | 49 | 18 | 91 | 202 | 228 | 264 | 317 | 310 | 420 |
| SEER Blacks, 1973–93[b] | 0 | 5 | 5 | 16 | 33 | 32 | 93 | 145 | 219 | 333 | 221 | 405 | 332 |
| **Melanomas—Skin** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 26 | 56 | 94 | 153 | 201 | 313 | 433 | 576 | 685 | 720 | 759 | 738 | 805 |
| Connecticut Whites, 1973–93[b] | 26 | 55 | 93 | 153 | 199 | 316 | 418 | 571 | 661 | 738 | 768 | 729 | 858 |
| SEER Whites, 1973–93[b] | 30 | 63 | 106 | 170 | 240 | 295 | 376 | 479 | 583 | 633 | 689 | 627 | 669 |
| **Non-Hodgkin's Lymphomas** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 36 | 39 | 50 | 71 | 119 | 182 | 228 | 349 | 418 | 470 | 579 | 514 | 568 |
| Connecticut Whites, 1973–93[b] | 37 | 39 | 46 | 70 | 118 | 181 | 231 | 362 | 442 | 484 | 610 | 511 | 572 |
| SEER Whites, 1973–93[b] | 29 | 35 | 47 | 70 | 107 | 172 | 215 | 303 | 399 | 485 | 554 | 573 | 523 |
| Connecticut Blacks, 1940–93 | 18 | 29 | 42 | 96 | 70 | 89 | 68 | 283 | 195 | 375 | 79 | 155 | 420 |
| SEER Blacks, 1973–93[b] | 15 | 39 | 42 | 76 | 70 | 109 | 136 | 178 | 160 | 226 | 245 | 194 | 269 |
| **Lung and Bronchus** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 2 | 4 | 8 | 20 | 66 | 154 | 301 | 584 | 855 | 1012 | 1167 | 1156 | 867 |
| Connecticut Whites, 1973–93[b] | 2 | 4 | 8 | 21 | 66 | 157 | 308 | 613 | 892 | 1067 | 1237 | 1204 | 851 |
| SEER Whites, 1973–93[b] | 2 | 3 | 8 | 22 | 56 | 141 | 283 | 566 | 895 | 1065 | 1177 | 997 | 768 |
| Connecticut Blacks, 1940–93 | 0 | 20 | 13 | 56 | 11 | 373 | 443 | 663 | 1113 | 1351 | 1164 | 802 | 455 |
| SEER Blacks, 1973–93[b] | 4 | 4 | 12 | 45 | 100 | 269 | 555 | 782 | 1193 | 1301 | 1303 | 1104 | 667 |
| **Colon/Rectum** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 8 | 9 | 24 | 42 | 132 | 273 | 548 | 1117 | 2036 | 2728 | 3856 | 4865 | 6001 |
| Connecticut Whites, 1973–93[b] | 8 | 8 | 24 | 43 | 130 | 276 | 553 | 1141 | 2083 | 2810 | 3867 | 4841 | 5925 |
| SEER Whites, 1973–93[b] | 6 | 13 | 23 | 48 | 101 | 246 | 522 | 943 | 1709 | 2444 | 3300 | 4158 | 4819 |
| Connecticut Blacks, 1940–93 | 8 | 13 | 29 | 57 | 177 | 333 | 570 | 1188 | 1507 | 2015 | 2041 | 3229 | 4798 |
| SEER Blacks, 1973–93[b] | 8 | 8 | 27 | 73 | 145 | 317 | 540 | 925 | 1411 | 2216 | 2531 | 2852 | 2907 |
| **Prostate** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 0 | 1 | 1 | 8 | 27 | 152 | 593 | 1608 | 3415 | 5643 | 7563 | 9147 | 10 147 |
| Connecticut Whites, 1973–93[b] | 0 | 1 | 1 | 8 | 27 | 152 | 593 | 1608 | 3412 | 5622 | 7512 | 9109 | 10 195 |
| SEER Whites, 1973–93[b] | 0 | 1 | 0 | 7 | 39 | 220 | 749 | 2041 | 4319 | 7245 | 9725 | 11 056 | 12 143 |
| Connecticut Blacks, 1940–93 | 8 | 0 | 0 | 22 | 41 | 370 | 1071 | 2377 | 5166 | 8001 | 7762 | 11 193 | 10 401 |
| SEER Blacks, 1973–93[b] | 0 | 1 | 1 | 14 | 58 | 384 | 1260 | 2906 | 5689 | 9474 | 10 945 | 13 090 | 12 200 |

[a] Prevalence proportions estimated using the Feldman *et al*. method.

[b] Estimates were adjusted using the corresponding modelled completeness index.

more specific survival functions for the various cancer sites, the modelled and empirical completeness indices were fairly close except possibly for certain cancers among women (i.e. colon, melanomas, breast, cervix, and all cancers combined). Using more complex survival functions is an area for further research. We illustrated that incompleteness due to limited length of follow-up is a major problem in estimating prevalence in most cancer registry areas. The majority of cancers considered required a longer registration period than 21 years to avoid underestimation bias. While the registration period was almost sufficient for prostate cancer and explained about 90% or better of all prevalent cases for stomach, lung and bronchus, and non-Hodgkin's lymphomas, it was insufficient for breast, colon/rectum, melanomas of the skin and, to a much larger extent, cervical cancer.

Cancer prevalence, which reflects in a single measure the effects of incidence and survival, is an important indicator of the burden of this disease in the population and on the health care system. Currently the Connecticut Tumor Registry is the

**Table 3** Cancer Site and Age Group Specific Prevalence Proportion Estimates for Women per 100 000 on 1 January 1994[a]

| | Age Group | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cancer Site (Invasive)** | 25–29 | 30–34 | 35–39 | 40–44 | 45–49 | 50–54 | 55–59 | 60–64 | 65–69 | 70–74 | 75–79 | 80–84 | 85–89 |
| **All Cancers** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 447 | 667 | 1089 | 1858 | 3054 | 4614 | 6429 | 9014 | 11 349 | 13 609 | 14 847 | 15 696 | 16 343 |
| Connecticut Whites, 1973–93[b] | 495 | 736 | 1175 | 1941 | 3174 | 4779 | 6722 | 9419 | 11 745 | 14 111 | 15 577 | 16 773 | 17 738 |
| SEER Whites, 1973–93[b] | 454 | 732 | 1168 | 1858 | 3046 | 4485 | 6291 | 8650 | 11 576 | 14 264 | 16 337 | 17 406 | 18 383 |
| Connecticut Blacks, 1940–93 | 224 | 359 | 830 | 1434 | 2429 | 3632 | 4938 | 6317 | 8142 | 9557 | 10 069 | 11 233 | 11 742 |
| SEER Blacks, 1973–93[b] | 242 | 430 | 750 | 1466 | 2540 | 3508 | 4897 | 6068 | 7990 | 9301 | 10 399 | 11 589 | 12 519 |
| **Stomach** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 2 | 2 | 2 | 5 | 4 | 14 | 36 | 39 | 51 | 105 | 164 | 182 | 179 |
| Connecticut Whites, 1973–93[b] | 2 | 2 | 2 | 5 | 4 | 14 | 38 | 40 | 51 | 104 | 157 | 184 | 184 |
| SEER Whites, 1973–93[b] | 1 | 1 | 3 | 3 | 6 | 13 | 26 | 32 | 44 | 70 | 100 | 130 | 148 |
| Connecticut Blacks, 1940–93 | 3 | 0 | 16 | 6 | 23 | 18 | 35 | 66 | 76 | 178 | 49 | 223 | 149 |
| SEER Blacks, 1973–93[b] | 1 | 5 | 5 | 8 | 9 | 21 | 41 | 76 | 80 | 126 | 179 | 162 | 285 |
| **Melanomas—Skin** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 51 | 92 | 155 | 221 | 266 | 344 | 423 | 457 | 475 | 430 | 482 | 354 | 406 |
| Connecticut Whites, 1973–93[b] | 51 | 83 | 161 | 224 | 274 | 358 | 446 | 495 | 508 | 439 | 506 | 406 | 503 |
| SEER Whites, 1973–93[b] | 63 | 109 | 183 | 249 | 322 | 347 | 383 | 448 | 493 | 456 | 455 | 459 | 465 |
| **Non-Hodgkin's Lymphomas** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 22 | 22 | 36 | 50 | 95 | 140 | 188 | 276 | 389 | 427 | 521 | 468 | 352 |
| Connecticut Whites, 1973–93[b] | 22 | 22 | 35 | 50 | 94 | 144 | 192 | 287 | 412 | 444 | 551 | 477 | 370 |
| SEER Whites, 1973–93[b] | 20 | 21 | 29 | 46 | 80 | 126 | 172 | 258 | 358 | 426 | 492 | 485 | 438 |
| Connecticut Blacks, 1940–93 | 12 | 10 | 37 | 44 | 80 | 94 | 97 | 197 | 123 | 201 | 142 | 196 | 149 |
| SEER Blacks, 1973–93[b] | 13 | 13 | 17 | 31 | 56 | 89 | 125 | 170 | 151 | 251 | 217 | 212 | 225 |
| **Lung and Bronchus** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 1 | 8 | 9 | 33 | 79 | 156 | 306 | 471 | 592 | 644 | 684 | 528 | 359 |
| Connecticut Whites, 1973–93[b] | 1 | 8 | 9 | 33 | 80 | 159 | 312 | 487 | 612 | 667 | 727 | 572 | 380 |
| SEER Whites, 1973–93[b] | 3 | 6 | 9 | 22 | 59 | 134 | 270 | 417 | 590 | 681 | 676 | 519 | 358 |
| Connecticut Blacks, 1940–93 | 0 | 7 | 0 | 26 | 89 | 76 | 287 | 350 | 544 | 399 | 671 | 346 | 683 |
| SEER Blacks, 1973–93[b] | 2 | 7 | 13 | 24 | 96 | 120 | 274 | 387 | 590 | 487 | 487 | 333 | 309 |
| **Colon/Rectum** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 1 | 15 | 22 | 57 | 119 | 214 | 463 | 820 | 1282 | 2017 | 2668 | 3538 | 4326 |
| Connecticut Whites, 1973–93[b] | 1 | 14 | 20 | 52 | 112 | 209 | 476 | 856 | 1353 | 2100 | 2762 | 3628 | 4638 |
| SEER Whites, 1973–93[b] | 5 | 10 | 22 | 51 | 111 | 205 | 420 | 768 | 1226 | 1807 | 2466 | 3262 | 4050 |
| Connecticut Blacks, 1940–93 | 0 | 14 | 27 | 64 | 119 | 318 | 554 | 732 | 1361 | 1656 | 1687 | 2995 | 3470 |
| SEER Blacks, 1973–93[b] | 1 | 14 | 33 | 74 | 152 | 291 | 520 | 839 | 1215 | 1789 | 2148 | 2531 | 3040 |
| **Breast** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 27 | 82 | 281 | 699 | 1383 | 2193 | 2989 | 4167 | 5143 | 5953 | 6152 | 6527 | 6833 |
| Connecticut Whites, 1973–93[b] | 27 | 82 | 281 | 698 | 1380 | 2190 | 3002 | 4124 | 4990 | 5732 | 5978 | 6487 | 6905 |
| SEER Whites, 1973–93[b] | 21 | 84 | 272 | 627 | 1312 | 2091 | 2801 | 3796 | 4887 | 5784 | 6378 | 6655 | 6911 |
| Connecticut Blacks, 1940–93 | 7 | 119 | 354 | 676 | 1239 | 1896 | 2194 | 2834 | 3124 | 3585 | 3914 | 4793 | 4317 |
| SEER Blacks, 1973–93[b] | 22 | 115 | 256 | 641 | 1195 | 1619 | 2131 | 2445 | 3019 | 3446 | 3948 | 4181 | 4260 |
| **Cervix Uteri** | | | | | | | | | | | | | |
| Connecticut Whites, 1940–93 | 37 | 62 | 131 | 179 | 210 | 256 | 272 | 278 | 346 | 383 | 358 | 377 | 414 |
| Connecticut Whites, 1973–93[b] | 37 | 62 | 129 | 182 | 219 | 261 | 282 | 316 | 375 | 471 | 510 | 517 | 711 |
| SEER Whites, 1973–93[b] | 31 | 78 | 128 | 188 | 240 | 296 | 377 | 438 | 529 | 609 | 700 | 897 | 959 |
| Connecticut Blacks, 1940–93 | 25 | 47 | 108 | 232 | 279 | 512 | 569 | 655 | 785 | 1080 | 884 | 990 | 969 |
| SEER Blacks, 1973–93[b] | 23 | 58 | 111 | 220 | 306 | 452 | 547 | 710 | 934 | 1281 | 1524 | 2069 | 2573 |

[a] Prevalence proportions estimated using the Feldman *et al*. method.

[b] Estimates were adjusted using the corresponding modelled completeness index.

only source of data in the US which allows us to directly compute prevalence. However, prevalence in Connecticut does not necessarily mirror that in the SEER areas (as suggested by differences in incidence and survival rates)[11] or the total US. For example, the estimated number of prevalent cases of any cancer for white men ages 70–74 in the US on 1 January 1994 (obtained by multiplying our modelled prevalence proportion estimates by the average of the populations in 1993 and 1994 from the Bureau of the Census), is 495 225 in Connecticut and 526 433 in SEER. Based on the modelled prevalence proportions on 1 January 1994 and projections of the white male population from the Bureau of the Census middle series,[20] in the year 2020 the estimated number of prevalent cases is 829 840 in Connecticut and 882 134 in SEER. Hence, the burden of cancer for white men in the US appears to be potentially very different when based on Connecticut data versus SEER data. The aim of this study was to provide prevalence proportions which better reflect the US white and black populations.

Factors influencing the number of years of follow-up required before the registration period is sufficient to capture the majority of prevalent cases includes the age in which the disease is common and the lethality of the disease. For example, the registration period was essentially sufficient for prostate cancer because it primarily occurs in old age where the life expectancy is relatively short. The registration period was also almost sufficient for lung and bronchus cancers because of the short survival associated with these diseases. On the other hand, for cancer of the cervix uteri the relatively young age at diagnosis and good survival require many more years of follow-up to capture prevalence. In general, women needed more years of follow-up than men, and whites more years of follow-up than blacks. This is because of better survival in women than men, and in whites than blacks.

The empirical completeness indices for whites from Connecticut could have been used to correct the SEER-based prevalence estimates rather than the modelled completeness indices for whites from SEER. However, random variation in the empirical estimates, and uncertainty about the representativeness of the Connecticut-based completeness indices to SEER indicated the need for modelled completeness indices. Further, sparse data limited us from obtaining empirical completeness indices for blacks in Connecticut. Yet although the modelled completeness indices based on SEER data are more stable, and can be obtained for blacks, they require certain assumptions about the cure fraction and distribution function of survival. Hence, limitations exist for both approaches.

Modelled prevalence estimates for whites in Connecticut compared to SEER were higher for stomach, non-Hodgkin's lymphomas, and colon/rectal cancers, and lower for prostate and cervix uteri cancers. This may be explained by higher incidence rates in Connecticut than in SEER for the former set of cancers but lower incidence rates in Connecticut compared to SEER for the latter set of cancers.[10] Prevalence estimates for blacks could not be directly compared between Connecticut and SEER, but we would expect that they would be higher in those areas displaying higher incidence rates. The incidence rates between Connecticut and SEER vary greatly for blacks for certain cancers (all cancers combined, stomach, lung and bronchus, prostate, and cervix uteri).[10] Hence, the modelled prevalence estimates among blacks in SEER for these cancer sites would be different than in Connecticut and better reflect US prevalence.

While the modelled SEER-based prevalence estimates provide a better representation of US prevalence, methods to obtain US and state level prevalence estimates are of primary interest. This has led to prevalence estimates obtained from national surveys,[21] Medicare data,[22–24] and a recent effort based on general methods using mortality and survival data.[25] The National Cancer Institute is currently sponsoring a project developing and applying methods to obtain estimates of US incidence and prevalence. The estimates presented in the current work are important for validating estimates obtained in further modelling efforts.

## References


[1] Enstrom JE, Austin DE. Interpreting cancer survival rates. *Science* 1977;**195:**847.

[2] Feldman AR, Kessler L, Myers MH, Naughton MD. The prevalence of cancer. *N Engl J Med* 1986;**315:**1394–97.

[3] Adami OH, Gunnarsson T, Sparen P, Eklund G. The prevalence of cancer in Sweden 1984. *Acta Oncologia* 1989;**28:**463–70.

[4] Capocaccia R, Verdecchia A, Micheli A, Sant M, Gatta G, Berrino F. Breast cancer incidence and prevalence estimated from survival and mortality. *Cancer Causes Control* 1990;**1:**23–29.

[5] Polednak AP. Estimating prevalence of cancer in the United States. *Cancer* 1997;**80:**136–41.

[6] Teppo L, Hakama M, Hakulinen T, Lehtonen M, Saxen E. Cancer in Finland 1953–70: Incidence, mortality, prevalence. *Acta Pathologica Microbiologica Scandinavica (A)* 1975;**(Suppl.252)**.

[7] Micheli A, Gatta G, Sant M *et al.* Breast cancer prevalence measured by the Lombardy Cancer Registry. *Tumori* 1997;**83:**875–79.

[8] Connelly RR, Campbell PC, Eisenberg H. Central registry of cancer cases in Connecticut. *Public Health Rep* 1968;**83:**386–90.

[9] Gershman ST, Flannery JT, Barrett H, Nadel RK, Meigs JW. Development of the Connecticut Tumor Registry. *Conn Med* 1976;**40:**697–701.

[10] Ries LA, Kosary CL, Hankey BF, Miller BA, Harras A, Edwards BK (eds). *SEER Cancer Statistics Review, 1973–1994*. National Cancer Institute. NIH Pub. No. 97-2789. Bethesda, MD, 1997.

[11] Capocaccia R, De Angelis R. Estimating the completeness of prevalence based on cancer registry data. *Stat Med* 1997;**16:**425–40.

[12] Shryock HS, Siegel JS *et al.* US Bureau of the Census. *The Methods and Materials of Demography*. Third Print (rev.). Washington DC: US Government Printing Office, 1975.

[13] Ederer F, Axtell LM, Cutler SJ. The relative survival rate: a statistical methodology. *Natl Cancer Inst Monog* 1961;**6:**101–21.

[14] Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer* 1954;**8:**1–12.

[15] Cook PJ, Doll R, Fellingham SA. A mathematical model for the age distribution of cancer in man. *Int J Cancer* 1969;**4:**93–112.

[16] Goldman AI. Survivorship analysis when cure is a possibility: a Monte Carlo study. *Stat Med* 1984;**3:**153–63.

[17] Gamel JW, McLean IW, Rosenberg SH. Proportion cured and mean long survival time as functions of tumor size. *Stat Med* 1990;**9:**999–1006.

[18] Miller RG Jr. *Survival Analysis*. New York: John Wiley & Sons, 1981.

[19] *SEER 1973–93 Public-Use CD-Rom*. US Department of Health and Human Services, PHS/NIH/NCI/CSB, August 1996.

[20] Day JC. *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995–2050*. US Bureau of the Census, Current Population Reports, P25–1130, US Government Printing Office, Washington DC, 1996.

[21] Byrne J, Kessler LG, Devesa SS. The prevalence of cancer among adults in the United States: 1987. *Cancer* 1992;**68:**2154–59.

[22] McBean AM, Babish JD, Warren JL. Determination of lung cancer incidence in the elderly using Medicare claims data. *Am J Epidemiol* 1993;**137:**226–34.

[23] McBean AM, Warren JL, Babish JD. Measuring the incidence of cancer in elderly Americans using Medicare claims data. *Cancer* 1994; **73:**2417–25.

[24] Warren JL, Riley GF, McBean AM, Hakim R. Use of Medicare data to identify incident breast cancer cases. *Health Care Financing Rev* 1996; **18:**237–46.

[25] Verdecchia A, Capocaccia R, Egidi V, Golini A. A method for the estimation of chronic disease morbidity and trend from mortality data. *Stat Med* 1989;**8:**201–16.