

OVERVIEW

BACKGROUND AND DATA SOURCES

There are three measures that are commonly used to assess the impact of cancer in the general population. The **incidence rate** is the number of new cases per year per 100,000 persons. The **mortality rate** is the number of deaths per year per 100,000 persons. The **survival rate** is the proportion of patients alive at some point subsequent to the diagnosis of their cancer. All three measures are employed in this report. The Surveillance, Epidemiology, and End Results (**SEER**) Program (<http://seer.cancer.gov>) — based within the Surveillance Research Program (**SRP**) at the National Cancer Institute (**NCI**) — collects incidence and survival data for all areas that participate in the Program. The National Center for Health Statistics (**NCHS**) provides mortality data for the entire United States (**US**). All incidence and mortality rates in this report are age-adjusted (see below) to the 2000 US standard million population (see Appendix) unless otherwise specified. Age adjustment minimizes the effect of a difference in age distributions when comparing rates. Data are presented for a wide spectrum of cancers.

The annual *SEER Cancer Statistics Review* (**CSR**), containing the most recent incidence, mortality, and survival statistics, is published by the Cancer Statistics Branch of the NCI. The scope and purpose of the *CSR* follow a report to the Senate Appropriations Committee (Breslow, 1988), which recommended that a broad profile of cancer be presented regularly to the American public. This *CSR* includes incidence, mortality, and survival data from 1973 through 1999, the most recent year for which data are available. (Incidence data for 1999 may not be complete. Therefore, *exercise caution when comparing rates for 1999 with those for previous years.*) Most of the rates in this publication have been age-adjusted to the 2000 US standard million population. Previous publications have used the 1970 US standard million population. Therefore, this publication can not be compared to previous publications. This change conforms to a new federal policy for reporting disease rates and it allows for the age-adjusted rate to more reflect the current age distribution and burden of cancer.

Since 1996, the *CSR* has also been available (in .pdf format) at <http://seer.cancer.gov>. The website allows timelier distribution of the *CSR*. Additional SEER statistics can be obtained via **FastStats** or the **Cancer Query System**. The SEER public-use file with **SEER*Stat** software can be used over the internet or the user can order a CD-ROM. SEER*Stat provides a user-friendly PC desktop system for the production of a myriad of cancer statistics, such as incidence rates and survival rates, for various demographic and medical input variables. The SEER public-use data file contains information on over 3,200,000 tumors with no personal identifiers.

THE SEER PROGRAM

The National Cancer Act of 1971 mandated the collection, analysis, and dissemination of data useful in the prevention, diagnosis, and treatment of cancer. This mandate led to the establishment of the SEER Program. A continuing project of the NCI, the population-based cancer registries participating in the SEER Program routinely collect data on all cancers occurring in residents of the participating areas. Trends in cancer incidence and patient survival in the US are derived from this database.

The SEER Program is a sequel to two earlier NCI programs — the End Results Program and the Third National Cancer Survey. The initial SEER reporting areas were the States of Connecticut, Iowa, New Mexico, Utah, and Hawaii; the metropolitan areas of Detroit, Michigan, and San Francisco-Oakland, California; and the Commonwealth of Puerto Rico. Case ascertainment began with January 1, 1973, diagnoses.

In 1974-1975, the program was expanded to include the metropolitan area of New Orleans, Louisiana, the thirteen-county Seattle-Puget Sound area in the State of Washington, and the metropolitan area of Atlanta, Georgia. New Orleans participated in the program only through the 1977 data collection year. In 1978, ten predominantly black rural counties in Georgia were added. American Indian residents of Arizona were added in 1980. In 1983, four counties in New Jersey were added with coverage retrospective to 1979. New Jersey and Puerto Rico participated in the program until the end of the 1989 reporting year. The National Cancer Institute also began funding a cancer registry that, with technical assistance from SEER, collects information on cancer cases among Alaska Native populations residing in Alaska. In 1992, the SEER Program was expanded to increase coverage of minority populations, especially Hispanics, by adding Los Angeles County and four counties in the San Jose-Monterey area south of San Francisco. In 2002, the SEER Program expanded coverage to include Kentucky and Greater California (the counties of California that were not already covered by SEER). Also in 2002, New Jersey and Louisiana became SEER participants again.

The long-term incidence trends and survival data for this report are from five states — Connecticut, Hawaii, Iowa, New Mexico, and Utah — and four metropolitan areas — Detroit, Atlanta, San Francisco-Oakland, and Seattle-Puget Sound (Fig. I-1). Additional tables, show more recent incidence rates and trends for SEER 12 areas (the 9 areas above plus Los Angeles, San Jose-Monterey, and the Alaska Native Registry) since 1992.

The participating regions were selected principally for their ability to operate and maintain a population-based cancer reporting system and for their epidemiologically significant population subgroups. With respect to selected demographic and epidemiologic factors, they are when combined a reasonably representative subset of the US population. Data from the 9 or 12 SEER geographic areas used in this report represent, respectively, approximately 10 or 14 percent of the US population. By the end of the 1999 diagnosis year, the database contained information on over 3,200,000 cases diagnosed since 1973. Over 170,000 new cases are added annually.

The goals of the SEER Program are:

- (1) to assemble and report, on a periodic basis, estimates of cancer incidence and mortality in the US;
- (2) to monitor annual cancer incidence trends to identify unusual changes in specific forms of cancer occurring in population subgroups defined by geographic and demographic characteristics;
- (3) to provide continuing information on trends over time in the extent of disease at diagnosis, trends in therapy, and associated changes in patient survival; and
- (4) to promote studies designed to identify factors amenable to cancer control interventions, such as:
 - (a) environmental, occupational, socioeconomic, dietary, and health-related exposures;
 - (b) screening practices, early detection and treatment; and
 - (c) determinants of the length and quality of patient survival.

Incidence and survival data: The SEER Program contracts with nonprofit, medically-oriented organizations having statutory responsibility for registering diagnoses of cancer among residents of their respective geographic coverage areas. Each SEER contractor:

- (1) maintains a cancer information reporting system;
- (2) abstracts records for *resident* cancer patients seen in every hospital both inside and outside the coverage area;
- (3) abstracts all death certificates of *residents* (dying both inside and outside the coverage area) on which cancer is listed as a cause of death;
- (4) ensures complete ascertainment of cases by searching records of private laboratories, radiotherapy units, nursing homes, and other health services units that provide diagnostic service;
- (5) registers all in situ and malignant neoplasms (with the exceptions of certain histologies for cancer of the skin and (beginning in 1996) in situ neoplasms of the cervix uteri);

- (6) records data on all newly diagnosed cancers, including selected patient demographics, primary site, morphology, diagnostic confirmation, extent of disease, and first course of cancer-directed therapy;
- (7) provides active follow-up on all living patients (except for those with in situ cancer of the cervix uteri);
- (8) maintains confidentiality of patient records;
- (9) semiannually submits electronically to NCI data on all reportable diagnoses of cancer made in residents of the coverage area.

For 1992 to 2000 diagnoses, the SEER program codes site and histology by the *International Classification of Diseases for Oncology*, second edition (**ICD-O-2**) (Percy, Van Holten, & Muir, 1990). All cases before 1992 were machine-converted to ICD-O-2. Beginning with 2001 diagnoses, cases will be coded according to the third edition (**ICD-O-3**) (Fritz et al., 2000). The primary site groupings used for incidence are found in the Appendix. Follow-up rates are also in the Appendix.

Mortality data: The SEER Program annually obtains from the NCHS a public-use file containing information on all deaths occurring in the US by calendar year. Information on each death includes age at death, sex, geographic area of residence, and underlying and contributing causes of death. For this publication, only the underlying cause of death is used in the calculation of mortality rates. Cause of death before 1999 was coded according to ICD-9; beginning with deaths in 1999, ICD-10 was used. Mortality rates for the SEER geographic areas, for each state, and for the entire US are obtained from these data. A list of the mortality site groupings used in this publication is in the Appendix.

Numbers of estimated cancers and deaths in 2002: The SEER Program has obtained from the American Cancer Society (**ACS**) projections of the numbers of cancer cases and cancer deaths in the US in 2002. The ACS projects incidence in 2002 based on incidence rates from SEER for 1979-98 to the 2002 estimated US population (Jemal et al., 2002).

Population data: The county population estimates implemented in the NCI SEER*Stat software for the calculation of cancer incidence and mortality rates are available at <http://seer.cancer.gov/popdata/download.html>. They represent a modification of the annual time series of July 1 county population estimates (by age, sex, race, and Hispanic origin) produced by the Population Estimates Program of the Bureau of the Census with support from the NCI through an interagency agreement. The Census Bureau's population estimates are available on their Web site, <http://eire.census.gov/popest/data/counties.php>. The procedures implemented by the Census Bureau in producing their county estimates are documented at http://www.census.gov/population/estimates/county/casrh_doc.txt. Please refer to <http://seer.cancer.gov/popdata/methods.pdf> for specific documentation regarding modifications made by the NCI to the Census Bureau estimates. The following summarizes these modifications.

The initial modification affects only population estimates for the State of Hawaii. The Epidemiology Program of the Hawaii Cancer Research Center has developed its own set of population estimates, based on sample survey data collected by the Hawaii Department of Health.

This effort grew out of a concern that the native Hawaiian population has been vastly undercounted in previous censuses. The "Hawaii-adjustment" to the BOC estimates has the net result of reducing the estimated white population and increasing the Asian and Pacific Islander population for the state. The Bureau of Census estimates for the total population, black population, and American Indian and Alaska Native populations in Hawaii are unaffected.

An additional modification to the Census Bureau estimates was recently implemented in order to meet requirements for age-adjustment according to the year 2000 US population standard. Population estimates for two new age groups (less-than-one-year-olds and one-to-four-year-olds) were created for the 1969-89 estimates and for the 1990-99 estimates.

DEFINITIONS

Several technical terms are used in presenting the data in this report. Their definitions are presented here to clarify them for the reader.

Incidence rate: The cancer incidence rate is the number of new cancers of a specific site/type occurring in a specified population during a year, usually expressed as the number of cancers per 100,000 population at risk. That is,

$$\text{Incidence rate} = (\text{New cancers} / \text{Population}) \times 100,000.$$

The *numerator* of the incidence rate is the number of new cancers; the *denominator* of the incidence rate is the size of the population. The number of new cancers may include multiple primary cancers occurring in one patient. *The population used depends on the rate to be calculated.* For cancer sites that occur in only one sex, the sex-specific population (e.g., females for cervical cancer) is used.

The incidence rate can be computed for a given type of cancer or for all cancers combined. Except for 5-year age-specific rates, all incidence rates are *age-adjusted* (see below) to the 2000 US standard population (or, where appropriate, to the world standard million population). (In previous editions of the *CSR*, the 1970 US standard million population was used; therefore, incidence rates in this edition can not be compared to those in previous editions.) Incidence rates are for *invasive cancer only*, unless otherwise specified. (An exception is the incidence rate for cancer of the urinary bladder; there both in situ and invasive cancers are counted.)

Mortality rate: The cancer mortality rate is the number of deaths with cancer given as the underlying cause of death occurring in a specified population during a year, usually expressed as the number of deaths due to cancer per 100,000 population. That is,

$$\text{Mortality Rate} = (\text{Cancer Deaths} / \text{Population}) \times 100,000.$$

The *numerator* of the mortality rate is the number of deaths; the *denominator* of the mortality rate is the size of the population. As with the incidence rate, *the population used depends on the rate to be calculated.* The mortality rate can be computed for a given cancer site or for all cancers combined. Except for 5-year age-specific rates, all mortality rates are *age-adjusted* (see below) to the 2000 US standard million population (or, where appropriate, to the world standard million population). (In previous editions of the *CSR*, the 1970 US standard million population was used; therefore, mortality rates in this edition can not be compared to those in previous editions.)

Standard million population: A **standard population** for a geographic area, such as the US or the world, is a table giving the proportions of the population falling into the age groups 0, 1-4, 5-9, ..., 80-84, and 85+. A **standard million population** for a geographic area is a table giving the number of persons in each age group 0, 1-4, ..., 85+ out of a theoretical cohort of 1,000,000 persons that is distributed by age in the same proportions as the population. Table A-7 shows the US 2000 and world standard million populations. This publication does not use the new world standard million population used in some World Health Organization mortality publications.

Age-adjusted rate: An age-adjusted incidence or mortality rate is a weighted average of the age-specific incidence or mortality rates, where the weights are the proportions of persons in the corresponding age groups of a standard million population. The potential confounding effect of age is reduced when comparing age-adjusted rates computed using the same standard million population. For this report, the 2000 US standard million population and the world standard million population are used as the weights in computing age-adjusted rates, unless otherwise noted.

Percent change: The percent change (PC) in a statistic over a given time interval is
$$\text{Percent change} = (\text{Final value} - \text{Initial value}) / \text{Initial value} * 100.$$
(In this report, the initial year is usually 1973, and the final year is usually 1999.) A positive PC corresponds to an increasing trend, a negative PC to a decreasing trend.

Estimated annual percent change: The estimated annual percent change (EAPC) is calculated by first fitting a regression line to the natural logarithm of the rates (r) using calendar year (x) as a regressor variable. In this report we use the method of *weighted least squares* to calculate the regression equation. If $\ln(r)=mx+b$ is the resulting regression equation (with slope m), then $EAPC = 100(e^m - 1)$. A positive EAPC corresponds to an increasing trend, a negative EAPC to a decreasing trend.

Because the methods used in their calculation are mathematically different, *the signs of the PC and the EAPC for a given statistic and time interval may differ*, as occurs in a few of the tables presented. That is, one of these statistics may show an increasing trend, the other a decreasing trend.

Testing the hypothesis that the actual mean annual percent change is 0 is equivalent to testing the hypothesis that the theoretical slope estimated by the slope m of the line representing the equation $\ln(r)=mx+b$ is 0. The latter hypothesis is tested using the t distribution of m/SE_m with $n-2$ degrees of freedom. The standard error of m , SE_m , is obtained from the fit of the regression (Kleinbaum et al., 1988). (This calculation assumes that the rates increased or decreased at a constant rate over the entire calendar year interval; the validity of this assumption was not assessed.) In those few instances where at least one of the rates was 0, the linear regression was not calculated.

Life table: A table for a given population listing, for each sex and each age from 0 to 120, how many members die at that age and how many survive one more year.

Observed survival rate: The observed survival rate represents the proportion of cancer patients surviving for a specified time interval after diagnosis. Note that some of those not surviving died of the given cancer and some died of other causes.

Relative survival rate: The relative survival rate is calculated using a procedure (Ederer et al., 1961) whereby the observed survival rate is adjusted for expected mortality. The relative survival rate approximates the likelihood that a patient will not die from causes associated specifically with the given cancer before some specified time after diagnosis. It is always larger than the observed survival rate for the same group of patients.

Standard error: The standard error of a rate is a measure of the sampling variability of the rate.

Person-years of life lost: The person-years of life lost (PYLL) is calculated as follows: For each individual who dies of the cancer of interest, the number of years of expected additional life for an average person of that age and sex is obtained from life tables for the US population (available from the NCHS). The PYLL in the general population associated with a particular cancer is simply the sum of this expectation over all those individuals who died of that cancer in a particular year.

Average years of life lost: The average years of life lost (AYLL) associated with a particular cancer is the PYLL associated with that cancer in the general population divided by the number of deaths from that cancer in the general population in a specific year.

Prevalence: Prevalence is defined as the number or percent of people alive on a certain date in a population who previously had a diagnosis of the disease. It includes new (incidence) and pre-existing cases and is a function of both past incidence and survival.

Stage of disease at diagnosis: Extent of disease information determines stage of disease at diagnosis. The historical stage presented has four levels. An invasive neoplasm confined entirely to the organ of origin is said to be **localized**. A neoplasm that has extended beyond the limits of the organ of origin, either directly into surrounding organs or tissues or into regional lymph nodes, is said to be **regional**. A neoplasm that has spread to parts of the body remote from the primary tumor, either by direct extension or by discontinuous metastasis, is said to be **distant**. When information is not sufficient to assign a stage, a neoplasm is said to be **unstaged**. In situ tumors (except those of the cervix uteri) are also collected by SEER but generally are not published in this series. For some cancers and diagnosis years, the extent of disease information can also be converted to Stages 0-IV as defined by the American Joint Committee on Cancer (Beahrs et al., 1988).

SUMMARY TABLES

While there are detailed tables in separate sections for each of the major cancer sites, information on some rare cancers can be found in the summary tables of section I. For a detailed list of primary sites, the summary tables provide incidence and mortality rates for the most recent 5-year period, trends — percent change (PC) and estimated annual percent change (EAPC) — from 1973 to the most recent year, median age at diagnosis, median age at death, and survival rates. The information is provided by race (all races, whites, blacks) and by sex.

LONG-TERM TRENDS, 1950-1999

Trends in cancer mortality from 1950 to 1999 are summarized by age both for all cancers combined (Table I-2). These mortality data are based on experience in the entire US.

Summaries of long-term trends in cancer incidence, mortality, and survival are outlined in Table I-3. The first two columns of the table show the estimated number of new cancers and the reported number of cancer deaths for 1999; the next four columns show incidence and mortality changes from 1950 to 1999. Both the percent change (PC) and the estimated annual percent change (EAPC) for incidence are based on data from the five geographic areas for which data are available for each of three time periods: around 1950, 1969-71, and 1973-99. Due to the limited availability of incidence data from the early time periods and the change in the composition of the nonwhite population over time, the incidence trends are presented for whites only. The estimates for children are for children of all races in Connecticut only. Mortality data are for the entire US; they are for whites only in order to be comparable to the incidence data. The last two columns display 5-year relative survival figures for patients diagnosed during two time intervals, 1950-54 and a recent time period; the figures are based on information from the End Results program for 1950-54 and SEER for the recent time period.

Use caution when interpreting these statistics. Evaluating trends over a long period of time may hide recent changes in the trends.

YEARS OF LIFE LOST DUE TO PREMATURE DEATH FROM VARIOUS CAUSES

Mortality rates alone give an incomplete picture of the burden deaths impose on the population. Another measure, which adds a different dimension, is the years of life lost due to premature death. This shows the extent to which life is cut short by a particular cause or disease.

This measure is estimated by linking life table data to each death of a person of given age and sex. The life table permits a determination of the number of additional years an average person of that age and sex would be expected to live. In this report, the age groups used in the calculation were 1-year intervals. These remaining years of life left are summed over all deaths due to a particular cause, yielding the estimate of the number of person-years of life lost (PYLL). The average years of life lost (AYLL) is obtained by dividing the PYLL by the number of deaths. Both of these measures can be calculated for any cause of death.

CANCER PREVALENCE

Prevalence is defined as the number or percent of people alive on a certain date in a population who previously had a diagnosis of the disease. It includes new (incidence) and pre-existing cases and is a function of both past incidence and survival. Tables on prevalence are not shown in the overview this year. Instead, there is a special section devoted to prevalence which includes prevalence estimates and a discussion of statistical methods used to calculate prevalence.

PROBABILITY OF BEING DIAGNOSED WITH OR DYING FROM CANCER

Each site-specific section of the book contains a table showing the probability (expressed as a percent) of a person of a specified race, sex, and age (0, 10, 20, 30, 40, 50, or 60) being diagnosed with the specified cancer within the next 10, 20, or 30 years or within their remaining lifetime. Lifetime risks of being diagnosed with cancer and lifetime risks of dying from cancer also appear (as percents) in each table. There are summary tables of lifetime risk in the overview.

Lifetime and interval risks of being diagnosed with cancer: The probability of being diagnosed with cancer is computed by applying cross-sectional age-specific 1997-99 incidence rates from the 12 SEER areas and mortality rates from the entire US to a hypothetical cohort of 10,000,000 live births. This cohort is considered to be at risk for two mutually exclusive events: (1) developing the specified cancer, and (2) dying of other causes without developing the specified cancer. Using these two types of events, a standard **multiple decrement life table** (with 20 age groups from 0-4 to 90-94 and 95+) is derived. For each age interval, the number alive and free of the specified cancer at the beginning of the interval is decremented by the number who develop the specified cancer and the number who die of other causes. The lifetime risk of being diagnosed with the specified cancer is derived by summing all cancer cases from age 0-4 through age 95+ and dividing by 10,000,000. This calculation does not assume an individual lives to any particular age; rather, it is the sum over all age intervals of the probability of living to the beginning of that interval without developing the given cancer times the probability of developing the cancer in that interval. The probability of developing cancer during any time period (e.g., within 10 years of a specific age say 50) is calculated by adding up all the cancers in the life table over the specified age range and dividing by the number of individuals alive and free of the specified cancer at the beginning of the period (Feuer et al., 1992; Feuer et al., 1993). To improve the precision of the calculations, rates were calculated for the age groups 85-89, 90-94, and 95+. The BOC provided populations for these age groups for 1990 to 1999.

Lifetime risk of dying from cancer: The lifetime risk of dying from a specified cancer is derived using a standard multiple decrement life table (Elandt-Johnson & Johnson, 1980). For each age, the risks of dying of the specified cancer and of all other causes are calculated, based on mortality data from the entire United States.

U.S. CANCER MORTALITY RATES BY STATE

Each cancer-site-specific section of the book presents, for all 50 states and the District of Columbia, average annual mortality rates for the most recent 5-year period for all persons, males only, and females only. The rates are per 100,000 persons; they are age-adjusted to the US 2000 standard million population. (In previous editions of the CSR, the 1970 US standard million population was used; therefore, mortality rates in this edition differ from those in earlier editions.)

The five states with the highest rates and the five states with the lowest rates are identified. The states are also ranked from highest rate to lowest rate for each of the cancers for which rates are reported.

The **percent difference (PD)** between a state rate and the rate for the total US is given by the formula:

$$PD = [(State\ rate - Total\ US\ rate) / Total\ US\ rate] \times 100.$$

The **standard error** for each age-adjusted state rate is calculated, based on the assumptions that (1) for each age-specific rate, the number of deaths is a Poisson random variable (Keyfitz, 1966) and (2) the variance of the age-adjusted rate is a linear combination of the variances of the age-specific rates (Snedecor & Cochran, 1980; pp. 188-9).

The **standard error of the difference** (SE_d) between a state rate and the total US rate is given by the formula

$$SE_d = \sqrt{SE_s^2 + SE_u^2}$$

where SE_s and SE_u are the standard errors of a state rate and of the total US rate, respectively. The variance of each rate (i.e., the square of the standard error) is based on the Poisson assumption. The standard error does not represent the total error that may be present in the age-adjusted rate; it is merely the square root of the variance associated with the rates. In addition to this variance, there also exist potential biases and errors in the measurement of the rate that are difficult to assess accurately and probably impact differently on the error calculations for different states.

The difference between each age-adjusted state rate and the age-adjusted US rate is tested for statistical significance (see below) by calculating a Z (standard normal) statistic from the formula:

$$Z = (\text{State rate} - \text{Total US rate}) / SE_d$$

Although the rates being compared are not independent because each state is part of the US, this doesn't substantially affect the statistical test because each state represents a small proportion of the total US.

SOCIOECONOMIC STATUS AND CANCER

As a supplement to this year's *SEER Cancer Statistics Review*, a publication entitled "**Area Socioeconomic Variations in Cancer Incidence, Mortality, Disease Stage, and Survival, 1975-1999**" will be released in the Summer of 2002. This publication includes analyses of trends and patterns in rates of incidence, mortality, stage of disease at diagnosis, and survival from all cancers combined and from lung, colorectal, prostate, breast, cervical, and melanoma cancers in relation to two area socioeconomic measures: median family income and percentage of population with at least a high school diploma. Cancer incidence and mortality trends are presented for the 1975-1999 time period. Trends in cancer survival and stage at diagnosis are presented for the 1988-1999 time period. Area socioeconomic measures are defined at both the county and census tract levels using decennial census data. Cancer incidence, disease stage, and survival data are drawn from the SEER database, whereas cancer mortality data for the entire United States come from the national mortality database maintained by the National Center for Health Statistics.

MEASUREMENT ERRORS

Errors in the estimation of death rates can occur in either the numerator (the number of reported deaths) or the denominator (the size of the population). One possible source of numerator error is underregistration of deaths. Although investigation by the National Center for Health Statistics indicates that over 99% of all deaths in the US are registered, little is known about the possible existence of any differences in death registration by geographic area, age, sex, or race.

Numerator error also can occur due to misclassifications, especially of race, ethnicity, or cause of death. Research indicates that, for infant mortality, misclassification is highest for races other than white or black (Hahn et al., 1992). The true extent of racial or ethnic misclassifications in death certificate coding remains unknown.

In coding overall cancer mortality, misclassifications of cause of death would occur when either the true cause of death was cancer while a cause other than cancer was coded, or vice versa. Even if a death is correctly attributed to cancer, the primary cancer may be incorrectly identified. It is already known, for example, that this is a problem with primary liver cancer (Percy, Ries, & Van Holten, 1990).

Denominator errors arise through under- and overenumeration in the decennial census, which is the basis of intercensal population estimates and population projections. To the extent that any over- or undercount is substantial and variable among subgroups or geographic areas, it may have important consequences on calculated death rate statistics. The effect of an *undercount* of population is that it decreases the denominator, leading to an *overestimate* of the rate. Conversely, an *overcount* of population would result in an *underestimate* of the rate.

In 1980, underenumeration varied by age group, with the greatest difference found for those 80 and older, who were undercounted by about 5% (US Bureau of the Census, 1986). All other age groups were either over- or undercounted by less than 3%. For race-sex-age groups, the coverage was lowest for black males aged 40-49, who were undercounted by 19%. It is thought that no improvement was achieved with the 1990 census; in some instances, underenumeration may have been worse than in 1980.

Any of these errors alters the count in either the numerator or the denominator, which in turn affects the calculated rate. Since the types of error encountered may differ by type of cancer, age group, race, sex, or even state, their impact is difficult to ascertain. *Use caution when dealing with those areas where potential problems may be present.*

STATISTICAL SIGNIFICANCE

Errors may be made in the determination of a given statistic. In order to test whether two groups (such as the populations of a state and the entire US) have the same or different *actual* rates, the *observed* rates for the groups are compared. Statisticians consider that a difference in observed rates can be explained by one of two hypotheses: (H_0) The actual rates are really the same, but the observed rates are different because of some combination of error-causing factors, or (H_1) the actual rates of the groups are really different. H_0 is called the **null hypothesis** (because it says there is *no* real difference); H_1 is called the **alternate hypothesis**. Typically, H_0 is rejected only if there is strong evidence in favor of H_1 . (Thus, if the observed rates are equal, we cannot reject H_0)

Using statistical theory, one can determine the distribution of the rate difference under the assumption that H_0 is true. Then values of the rate difference that are very unlikely to occur if H_0 is true are identified. More specifically, a small positive number, called **alpha** (α), is chosen; usually, α is 0.05 or 0.01. (Alpha is called the **significance level** of the hypothesis test.) One can then identify limits for the difference in rates such that, if H_0 is true, the probability of the difference being outside of those limits is α . If the observed difference is *outside* of these limits, then the observed result is *very unlikely* to happen if H_0 is true, so H_0 is rejected.

Another way of looking at the same process is to calculate, assuming H_0 is true, the probability that the observed difference or any greater observed difference would occur; this number is called the **P-value** of the observed result. If the P-value of a comparison is less than α — that is, the observed difference is *very unlikely* to happen if the null hypothesis is true — H_0 will be rejected. If the P-value of a test is greater than the significance level α , H_0 will not be rejected. When a difference in rates is sufficiently large to cause the null hypothesis to be rejected for a given value of α , it is called a **statistically significant** difference.

When a null hypothesis is rejected, there remains a small chance that a wrong decision has been made. If many statistical comparisons are done, even with $\alpha = 0.01$, the chance of making at least one wrong decision becomes a concern. In testing the differences between the total US rate and the rate for each state (or for the District of Columbia) for a given cancer, 51 statistical comparisons of the type described above are performed. Based on one of Bonferroni's inequalities (if there are n events and p_i is the probability of success in event i , then $P(\text{at least 1 success}) < p_1 + \dots + p_n$) (Snedecor & Cochran, 1980; p. 115-117), the significance level α for each individual comparison was set equal to $0.01/51 \approx 0.0002$. Thus, only individual-state-to-total-US comparisons with an associated P-value less than 0.0002 are

considered to be statistically significant. That is, a *very small* significance level α (0.0002) is used in order to minimize the total risk (0.01) of falsely deciding that some pair of equal rates are unequal. *Use caution in assessing statistically significant differences.* Population size has an important role in any calculation of statistical significance. Some states may have estimated rates that are very close to the estimated total US rate, but because of their large population, the difference between their estimated rate and the estimated total US rate is found to be statistically significant. In this case, the true state rate and the true US rate are almost certainly different, because the observed difference, though small, is nearly impossible if the null hypothesis (equal rates) is true. A small difference in rates, however, may have no practical importance.

On the other hand, some smaller states may have estimated rates that differ substantially from the estimated total US rate, but because of their relatively small population, the differences are found to be statistically nonsignificant. When this happens, if the true state rate and the true US rate were equal, the probability of obtaining a difference at least as large as what has been observed is greater than $\alpha=0.0002$. Therefore, *because the evidence against it isn't strong enough, the null hypothesis (equal rates) is not rejected.*

If the percent difference (PD) between the two rates is small, there may be some question about the importance of the difference. It is difficult to specify a minimally significant absolute PD, below which the difference would always be unimportant, because the observed PD will depend on the populations of the areas involved. It may be of value to consider the size of the PD between a state rate and the US rate in assessing the importance of a statistically significant difference. To further assist readers in interpreting the data, the tables are footnoted to indicate variations of more than 15% between a state rate and the US rate.

Comparing individual state rates with the US rate and assessing statistical significance is not an appropriate procedure for assessing geographic clustering of state rates. Identification of states which may represent regional clusters of high or low rates would require additional statistical and graphical analyses.

For a number of cancers, the District of Columbia has the highest mortality rates. *Use caution when comparing cancer rates for the District with those from the 50 states.* The District is an entirely urban area, whereas a state includes urban, suburban, and rural areas. Mortality rates for many cancers are higher in urban areas. Also, the District has a higher percentage of blacks (about two-thirds) than any state; their higher mortality rates for several types of cancer elevate the overall rate for the District.

JOINPOINT REGRESSION ANALYSIS OF CANCER TRENDS

A recent advance in the presentation of cancer trends is the use of joinpoint models (Kim et al., 2000). In past issues of the *Cancer Statistics Review*, certain time intervals (e.g., 1973-1996) were specified and the estimated annual percent changes (EAPC) were computed over those intervals. The choices of where to start and where to end an interval were arbitrary and sometimes did not give an accurate picture of the trend for a given cancer site. For example, the rates might be increasing and decreasing in different parts of the same interval. For some sites, increases occurred in the earlier years, followed by declines in more recent years.

To achieve greater descriptive accuracy, the computer now statistically finds the number and location of places where a trend changes. The point (in time) where a trend changes is called a **joinpoint**. Trends may change in different ways at a joinpoint: from up to down, from down to up, from up to up at a different rate, or from down to down at a different rate. In order to find the most accurate set of joinpoints with corresponding fixed-rate intervals, we use a **joinpoint regression model**. Joinpoint regression models on the natural logarithms of the rates describe the trends by a sequence of connected straight line segments. Adjacent segments are connected at a joinpoint. Each segment has an associated EAPC. On a logarithmic scale, the segments are linear.

Joinpoint analysis first assumes no joinpoints are needed to describe the data accurately — i.e., the trend over the entire interval 1973-1999 does not change. Joinpoints are added in turn if they are statistically significant. Thus, in the final model, each joinpoint (up to a maximum of three) represents a significant change in trend. Smoother polynomial models may provide a good fit overall, but are less sensitive to what is occurring at the ends of the data, especially for the most recent points.

A Windows-based program, *Joinpoint*, is freely available at <http://srab.cancer.gov/joinpoint/>; it accepts data from the SEER*Stat program. Further details on joinpoint regression may be found in a previous *CSR* (Ries et al., 2000) and in the cited reference.

INTERPRETATION OF CANCER STATISTICS

When reviewing the various cancer incidence, mortality, and survival statistics provided in this report, be aware that a number of factors may affect the interpretation of many of these statistics.

Survival rates for all cancers combined: The mix of cancers changes over time as the incidence of some cancers increases and the incidence of others decreases. Thus, in calculating the survival rate for all cancers combined, the proportions corresponding to the specific cancers will also change over time. Therefore, the overall cancer survival rate can fluctuate even when the survival rates for site-specific cancers remain unchanged. (While it is possible to adjust the survival rate for all cancers combined on the basis of the relative frequency of each specific cancer in some specified reference period, rates adjusted in this manner differ by only a small amount from unadjusted rates. In the future, such an adjustment may become more important if there are substantial changes in the incidence of various cancers.)

Early detection/screening: The improved earlier detection and diagnosis of cancers may produce an *increase* in both incidence rates and survival rates. These increases can occur as a result of the introduction of a new procedure to screen subgroups of the population for a specific cancer; they need not be related to whether use of the screening test results in a decrease in mortality from that cancer. As the proportion of cancers detected at screening increases, presumably as a result of increased screening of the population, patient survival rates will *increase*, because they are based on survival time *after diagnosis*. The interval between the time a cancer is diagnosed by a screening procedure and the time when the cancer would have been diagnosed in the absence of screening is called **lead-time** (Zelen, 1976). (Screening for breast cancer has been demonstrated to result in increased survival over and above that resulting from lead-time alone and to reduce breast cancer mortality. The benefit of screening is being studied for some other cancers.)

If a new screening procedure consistently detects cancer in a preinvasive phase, this may result in a *decrease* in survival rates for *invasive* cancer. In this case, **length-biased sampling** (Zelen, 1976) may be operating. Length-biased sampling would result in the preferential detection — in a *preinvasive* phase — of those cancers that would have had a relatively good prognosis had they progressed to invasive disease; these potentially invasive cancers would be systematically eliminated. If this occurs, the mix of cancers that are not detected at screening and progress to invasive may become less prognostically favorable, resulting in a *decrease* in survival rates for patients with invasive cancers. (Length-biased sampling may at least partially explain survival trends for cervical cancer. Other cancers possibly affected include breast, colon, rectum, and prostate.)

Changes in diagnostic criteria: Early detection of cancer resulting from either screening or earlier response to symptoms may result in the increasing diagnosis of small tumors that are not yet life-threatening. This may have the effect of raising the incidence and survival rates with little or no change in mortality rates. Breast, colon, prostate, cervix uteri, bladder, and skin (melanoma) are the cancer sites most likely to be affected.

Technological advances in diagnostic procedures: In this report, trends in survival by stage at

diagnosis are not presented for specific cancers; trends in stage distributions are presented rarely. However, it is possible to compare survival rates by stage and stage distributions given here with those for earlier time periods (as provided in previous reports or available from the SEER public-use data file). Thus, it is necessary to comment on the effect of technological advances on the diagnosis and staging of cancer.

The assignment of a given stage to a particular cancer may change over time due to advances in diagnostic technology. Introduction of new technology can give rise to a phenomenon known as **stage migration**. Stage migration occurs when diagnostic procedures change over time, resulting in an increase in the probability that a given cancer will be diagnosed in a *more advanced* stage. For example, certain distant metastases that would have been undetectable a few years ago can now be diagnosed by a computer tomography (CT) scan or by magnetic resonance imaging (MRI). Therefore, some patients who would have been diagnosed previously as having cancer in a *localized* or *regional* stage are now diagnosed as having cancer in a *distant* stage. The likely result would be to remove the worst survivors — those with previously undetected distant metastases — from the localized and regional categories and put them into the distant category. As a result, the stage-at-diagnosis distribution for a cancer may become less favorable over time, but the survival rates for each stage may improve: the early stage will *lose* cases that will survive *shorter* than those remaining in that category, while the advanced stage will *gain* cases that will survive *longer* than those already in that category. However, *overall survival would not change* (Feinstein et al., 1985). Stage migration is an important concept to understand when examining temporal trends in survival by stage at diagnosis as well as temporal trends in stage distributions; it could affect the analysis of virtually all solid tumors.

Evolution of stage classifications: Every few years, the American Joint Committee on Cancer produces a new cancer-staging manual (Beahrs, 1988). The evolution of such classifications reflects the identification of new prognostic factors that may influence choice of treatment. The SEER Program collects data on **extent of disease (EOD)** rather than stage; EOD is *more specific* than stage and usually determines stage, even when stage definitions change. Thus, SEER easily adapts to changes in stage definitions; moreover, trends in a newly redefined stage can usually be calculated.

For those cancers for which new prognostic variables are introduced into staging, so that previously collected EOD data cannot determine new stage categories, there can be problems in assessing trends in stage of disease. Only by reviewing the evolution of staging for a given cancer is it possible to determine what effect changes in stage definitions have had on stage-specific survival and on stage-at-diagnosis distributions. Stage migration (mentioned above) and EOD migration need also be taken into account. One reason for using the historical categories of *localized*, *regional*, and *distant* is that these categories have been fairly consistent over time.

Interpreting relative survival rates: The relative survival rate is the ratio of the observed survival rate to the expected survival rate for a given patient cohort. When the base population used in calculating the expected survival is similar to the cohort of cancer patients except for the latter's cancer experience, the relative survival rate approximates the cancer's cause-specific survival rate. The expected survival rate is based on mortality rates for the entire population, taking into account, as appropriate, the age, sex, race, and year of diagnosis of the patients. Assuming that the presence of cancer is the only factor that distinguishes the cancer patient cohort from the general population, the relative survival rate approximates the probability that a patient will *not* die of the diagnosed cancer within the given time interval.

A factor related to the risk of a cancer may also be related to the risk of dying from causes unrelated to the cancer. An example of such a factor is *smoking*. Smoking is a major risk factor for lung cancer; therefore, a cohort of lung cancer patients will contain a much higher proportion of smokers than does the general population. However, smoking is also a risk factor for other diseases, resulting in smokers having a shorter life expectancy than nonsmokers. Expected survival rates for lung cancer patients based on the general population will be unduly optimistic for this reason; they will result in relative rates that are *lower* than they should be.

The problem cannot be easily corrected because separate life tables for smokers and nonsmokers are not available. Amount of smoking (usually measured in pack-years) is clearly an important variable. The possibility that expected rates may not be appropriate for a given patient cohort should also be considered when examining relative survival rates for patients with cancers of the cervix uteri or breast, because the risk of these cancers has been associated with socioeconomic status (Baquet et al., 1991), which may be related to life expectancy.

Previous to the CSR for 1973-1996, the expected rate tables used were for 1970 and 1980; there were separate tables for whites, blacks, American Indians, Chinese, Japanese, Filipinos, white Hispanics, and Hawaiians. In updating the tables for 1990, several problems emerged. The US life tables are based on age, race, and sex information from death certificates. The information on race on the death certificate may not be accurate (Rosenberg et al., 1999). One reason is that funeral directors may inaccurately report race on a death certificate. Also, reported age at death, especially for those older than 85, may not be accurate because birth certificates were not issued with as much regularity in the early 1900s as they are today. Although race misclassification and age-at-death misreporting exist across all races, they may be more problematic for races other than white or black because of those races' smaller population sizes. Therefore, life tables were generated for 1970, 1980, and 1990 only for white, black, and other; these life tables were used to produce the relative survival rates in this book. There may be small variations among survival rates calculated in this CSR and those in CSRs prior to 1973-1996.

Comparison with other databases: The SEER data are obtained from population-based cancer registries covering about 14 percent of the US population. It is sometimes of interest to compare cancer statistics for SEER areas with those from other registries both in the US and worldwide. In making such comparisons, one must carefully consider the factors considered above for both data sources. In addition, one should assess all of the following: (1) completeness of case ascertainment, (2) rules used to determine multiple primaries, (3) follow-up, (4) rules used in assigning and coding cause of death, and (5) the sources and procedures used in obtaining population estimates. Depending on the rates being compared, there could be other confounding factors which should be considered. The same standard million population should be used for the age-adjustment of each group being compared.

It is sometimes interesting to compare survival data for cancer patients in SEER areas with data from clinical trials. *This must be done with great caution.* Survival data from clinical trials may have been obtained from a patient population that differs from that of SEER patients in prognostic factors for the given cancer; any survival comparisons would have to adjust for such differences. Also, it is necessary to verify that the methodology used in computing survival rates is the same for both data sources. Furthermore, clinical-trials patients may differ from SEER patients in characteristics that may be related to survival but are not recorded in either database. If this were true for a given cancer, it would not be possible to make valid comparisons of this type.

Errors in data collection: In the process of registering cancer patients, errors may be made in abstracting and coding the data, which includes demographic information, cancer site, histology, extent of disease, treatment, and patient survival. Quality control studies are periodically carried out to detect and correct this type of error, but no attempt is made to incorporate this source of error into the variance estimates of cancer rates reported here.

Comparison of this report with previous reports: It is important to note that most rates in this CSR were age-adjusted to the 2000 standard million US population; in the past, the 1970 standard million population was used. Therefore rates in this report can not be compared to rates and trends in previous reports.

The cancer registries that participate in the SEER Program submit data on all cancers diagnosed in their coverage areas to the NCI each year. Because of the dynamic nature of the registries' databases, *the reported number of new cancer cases in a particular race-sex-age-cancer category may change for a calendar year for which data have already been reported in a previous publication.* Additional cancer cases that were previously overlooked for a given diagnosis year may have been found and reported to the central registry. There may have been follow-back of cancers diagnosed by death certificate only; successful efforts to establish the dates of diagnosis for such patients will change the number of patients reported for a given diagnosis year. Code changes may occur when a patient dies; for example, information on race is generally available on the death certificate and may be used to update a previously unknown value. There may have been elimination of duplicate records for the same patient, often due to name changes or misspellings.

Thus, a recent report may have a different number of cases for a given diagnosis year than an earlier report, with resulting effects on incidence and possibly survival rates. Population estimates may also change from one report to another for some calendar years. This occurs because the NCI receives population estimates that are regularly updated by the BOC; for example, previous population estimates for years beginning with 1990 were replaced with new estimates from the BOC. Such changes may result in some differences between incidence and mortality rates for a given calendar period as published in different reports.

STANDARD ERRORS OF RATES

Survival rates: In the tables presenting survival rates, the magnitude of the standard error is given as a clue to the reliability of a given rate: the greater the standard error, the less reliable the rate. In addition, if there were fewer than 25 diagnoses in the first interval of the life table constructed to calculate survival, or if all cases became lost to follow-up within an interval, a valid survival rate could not be calculated, as is noted in the table footnotes.

The **standard error (SE)** of a relative survival rate is obtained as follows (Ederer et al., 1961):

$$SE(CR_t) = CR_t \cdot \sqrt{\frac{q_1}{e_1 - d_1} + \frac{q_2}{e_2 - d_2} + \dots + \frac{q_t}{e_t - d_t}}$$

where CR_t is the t -year relative survival rate, and for $i = 1, \dots, t$,
 q_i is the probability of dying in year i after diagnosis,
 e_i is the effective number of patients at risk in year i after diagnosis, and
 d_i is the number of deaths in year i after diagnosis.

Incidence and mortality rates: The standard errors of age-adjusted incidence and mortality rates are often not specified. However, the reader can approximate the SE of a particular incidence or mortality rate by the following formula for the SE of a crude incidence or mortality rate (Keyfitz, 1966):

$$SE(\text{rate}) \approx \text{rate} / \sqrt{\text{cancer cases or deaths}}$$

Appendix Tables A-1 and A-2 provide numbers of cancer diagnoses within SEER areas and numbers of deaths in the entire US, respectively, by race and sex for the most recent 5-year period. These can be used to obtain approximations of the standard errors for associated age-adjusted rates for the same time period using the above formula. To approximate the standard error of a rate for a single year, use the formula but substitute the cancer cases or deaths with the number of cancer cases or deaths divided by 5.

REFERENCES

- Baquet CR, Horm JW, Gibbs T, Greenwald P. Socioeconomic factors and cancer incidence among blacks and whites. *J Natl Cancer Inst* 1991; 83:551-557.
- Behrs OH, Henson DE, Hutter RV, Myers MH, editors. *Manual for Staging of Cancer*, 3rd ed. Philadelphia (PA): Lippincott; 1988.
- Breslow L (Chairman, Extramural Committee to Assess Measures of Progress Against Cancer). *Measurement of progress against cancer: Final report to the Senate Appropriations Committee*. Bethesda: National Cancer Institute; 1988.
- Ederer F, Axtell LM, Cutler SJ. The relative survival rate: A statistical methodology. *J Natl Cancer Inst Monogr* 1961; 6:101-121.
- Elandt-Johnson RC, Johnson NL. *Survival Models and Data Analysis*. New York (NY): Wiley; 1980.
- Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon: Stage migration and new diagnostic techniques as a source of misleading statistics for survival of cancer. *New Engl J Med* 1985; 312:1604-1608.
- Feldman AR, Kessler L, Myers M, Naughton MD. The prevalence of cancer: Estimates based on the Connecticut Tumor Registry. *New Engl J Med* 1986; 315:1394-1397.
- Feuer EJ, Wun L-M, Boring CC. Probability of developing cancer. In: Miller BA, Ries LAG, Hankey BF, Kosary CL, Edwards BK, editors. *Cancer Statistics Review: 1973-1989*, National Cancer Institute, NIH Pub. No. 92-2789, 1992. p. XXX.1-8.
- Feuer EJ, Wun L-M, Boring CC, Flanders WD, Timmel MJ, Tong T. The lifetime risk of developing breast cancer. *J Natl Cancer Inst* 1993; 85:892-897.
- Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin DM, Whelan S, editors. *International Classification of Diseases for Oncology*, 3rd ed. Geneva: World Health Organization; 2000.
- Hahn RA, Mulinare J, Teutsch SM. Inconsistencies in coding of race and ethnicity between birth and death in U.S. infants. *JAMA* 1992; 267:259-263.
- Jemal A, Thomas A, Murray T, Thun M. Cancer statistics, 2002. *CA Cancer J Clin* 2002; 52:23-47.
- Keyfitz N. Sampling variance of standardized mortality rates. *Hum Biol* 1966; 38:309-317.
- Kim H-J, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Stat Med* 2000; 19:335-351.
- Kleinbaum DG, Kupper LL, Muller KE. *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. Boston (MA): PWS-Kent, 1988. p. 266-268.
- Percy C, Ries LAG, Van Holten VD. The accuracy of liver cancer as the underlying cause of death on death certificates. *Public Health Rep* 1990; 105:361-368.
- Percy C, Van Holten V, Muir C, editors. *International Classification of Diseases for Oncology*, 2nd ed. Geneva: World Health Organization; 1990.
- Ries LAG, Eisner MP, Kosary CL, Hankey BF, Miller BA, Clegg LX, Edwards BK (eds). *SEER Cancer Statistics Review, 1973-1997*, National Cancer Institute. NIH Pub. No. 00-2789. Bethesda, MD, 2000.

Rosenberg HM, Maurer JD, Sorlie PD, Johnson NJ, MacDorman MF, Hoyert DL, Spitler JF, Scott C. Quality of Death Rates by Race and Hispanic Origin: A Summary of Current Research. Hyattsville (MD): National Center for Health Statistics; Vital and Health Statistics, Series 2, No. 128, 1999.

Snedecor GW, Cochran WG. Statistical Methods, 7th ed. Ames (IA): Iowa State University Press; 1980.

US Bureau of the Census. Current Population Reports; Series P-25 No. 985. Washington (DC): US Government Printing Office; 1986.

Zelen M. Theory of early detection of breast cancer in the general population. In: Heuson J-C, Matthei WH, Rozenzweig M, editors. Breast Cancer: Trends in Research and Treatment. New York (NY): Raven Press; 1976. p. 287-299.