O-1. BACKGROUND AND DATA SOURCES

There are four measures that are commonly used to assess the impact of a cancer in the general population. The **incidence rate** is the number of new cases per year per 100,000 persons. The **death** (or **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. The **prevalence count** is the number of people alive that have ever been diagnosed with a cancer. All four 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 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, prevalence, 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, prevalence, and survival data from 1975 through the most recent year for which data are available. Observed incidence data for the most recent years may not be complete. Therefore, delay adjusted rates are presented to compensate for this problem (see Reporting Delay).

While most of the rates in this publication have been age-adjusted to the 2000 US standard population, some previous SEER publications have used the 1970 US standard million population. Therefore, rates given in this publication cannot be compared to rates given in those publications. This change conforms to a new federal policy for reporting disease rates and it allows for the age-adjusted rate to more accurately reflect the current age distribution and burden of cancer.

Since 1996, the *CSR* has been available (in .pdf format) at *http://seer.cancer.gov*. This edition can be found at *http://seer.cancer.gov/csr/1975_2005/*. The website allows timelier distribution of the *CSR*. Additional SEER data can be obtained via ***FastStats*** (*http://seer.cancer.gov/faststats/*) or ***Cancer Query Systems***, an interactive system at *http://seer.cancer.gov/canques*, which allows the user to access over 10,000,000 cancer statistics. The SEER limited-use file with ***SEER*Stat*** software can be used over the internet, or the user can order a CD-ROM version at *http://seer.cancer.gov/data/options.html*. *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.

O-2. 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. The population-based cancer registries participating in NCI's 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 African-American counties in **rural 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 2001, the SEER Program expanded coverage to include **Kentucky, Greater California** (the counties of California that were not already covered by SEER), **New Jersey,** and **Louisiana**.

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); this set of registries is called the **SEER 9**. Additional tables show more recent incidence trends for the **SEER 13** areas (the 9 areas above plus Los Angeles, San Jose-Monterey, Alaska Native Registry, and rural Georgia) since 1992. Other tables give statistics for the **SEER 17** areas; these are the SEER 13 plus Kentucky, Greater California, New Jersey, and Louisiana.

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, 13, or 17

SEER geographic areas are used in this report; the given areas contain, respectively, approximately 9,14, or 26 percent of the US population. By the end of the 2005 diagnosis year, the database of 13 SEER and 4 expansion registries (plus Arizona Indians) contained information on over 7 million cases diagnosed since 1973. New cases added in the most recent data year (not including Arizona Indians) numbered over 370,000.

The goals of the SEER Program are:

(1) to assemble and report, on a periodic basis, estimates of cancer incidence, mortality, survival, and prevalence 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.

O-3. 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) strives for 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) at least annually 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 have been 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. Changes were made to the site recode for ICD-O-2 for comparability with cases coded to ICD-O-3. Follow-up rates are also in the Appendix.

A recent policy change of the Department of Veterans Affairs (VA) regarding sharing of VA cancer data has resulted in incomplete reporting of VA hospital cases in some central cancer registries for the 2005 data year. The section on VA reporting quantifies the missing number of VA patients in the SEER registries and provides adjustments of new case counts for 2005 based on prior years information. These VA adjustment factors may be used to correct for underreporting of 2005 age-specific incidence rates or age-adjusted incidence rates for SEER-9 and SEER-17 regions.

Excluded cancers: Some cancers were excluded from most of the analyses. Myelodysplastic syndrome (MDS), for example, was reclassified in ICD-O-3 (effective diagnosis year 2001) from nonmalignant to malignant; other cancers so reclassified include endometrial stromal sarcoma (low grade), papillary ependymoma, papillary meningioma, polycythemia vera, chronic myeloproliferative disease (NOS), myelosclerosis with myeloid metaplasia, essential thrombocythemia, refractory anemia, refractory anemia with sideroblasts, refractory anemia with excess blasts, and refractory anemia with excess blasts in transformation. In contrast, borderline tumors of the ovary were reclassified from malignant to nonmalignant at the same time. In addition, benign brain/CNS tumors were collected beginning for 2004 diagnoses. All of these cancers were excluded from most of the analyses, especially time trends. Pilocytic astrocytoma, although reclassified in ICD-O-3, was not excluded. Separate tables for MDS and benign brain/CNS are shown.

O-4. MORTALITY DATA

The SEER Program annually obtains from the National Center for Health Statistics (NCHS) a 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 death rates. Cause of death for 1969-1978 was coded according to ICD-8; for 1979-1998, ICD-9 was used; 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 and reflects updates made in 2004.

O-5. POPULATION DATA

The population estimates used in the SEER*Stat software to calculate cancer incidence and mortality rates for this report are a modified version of the annual time series of July 1 county population estimates by age, sex, race, and Hispanic origin that are produced by the Population Estimates Program of the US Census Bureau (*http://www.census.gov/popest/estimates.php*) with support from the NCI through an interagency agreement. Descriptions of the methodologies employed by the Census Bureau for various sets of estimates may be found on the same website. County population estimates for 2000 and later years must be bridged from 31 race categories used in Census 2000 to the four race categories specified under earlier OMB standards in order to report long-term cancer trends. The bridging methodology was developed by the National Center for Health Statistics and is described in a report (Ingram et al., 2003) and on their website *http://www.cdc.gov/nchs/about/major/dvs/popbridge/popbridge.htm*.

Modifications made by the NCI to the population estimates are documented in "Population Estimates Used in NCI's SEER*Stat Software" (*http://seer.cancer.gov/popdata/methods.html*) and the population data files are available for download (see "Download US Population Data" from *http://seer.cancer.gov/popdata/download.html*). Several of the modifications pertaining to the grouping of specific counties needed to assure the compatibility of all incidence, mortality and population datasets. Another 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 Census Bureau's estimates has the net result of reducing the estimated white population and increasing the estimated Asian and Pacific Islander population for the state. The estimates for the total population, black population, and American Indian and Alaska Native populations in Hawaii are not modified.

The 2001-2005 cancer incidence and mortality rates for American Indians and Alaska Natives (AI/AN) are based on the geographic areas (counties) included in the Indian Health Service's Contract Health Service Delivery Area (CHSDA). This reflects a concern that previously reported AI/AN rates were underestimated due to racial/ethnic misclassification of American Indian cases in geographic areas outside of CHSDA. This change has the net effect of higher, and more accurate, incidence and mortality rates for this population.

Usually the use of a population estimate for July 1 of a particular year reflects the average population of that area for the year. Both Hurricane Katrina and Hurricane Rita struck the Gulf Coast area of the United States in 2005. This had the effect of displacing large populations. Since there weren't any population estimates by age, race, sex, and county for time periods just after the hurricanes, it is very difficult to estimate the actual population at risk for certain areas along the Gulf Coast for 2005. For Louisiana, only the first six months of incidence data for

2005 coupled with ½ of the population estimate for July 1, 2005, were used to calculate cancer incidence. For death rate calculations, no adjustments were made to the total U.S. population, but for the Gulf area, an adjustment for displaced populations was made for 2005 state rates. For more details, see http://seer.cancer.gov/popdata/methods.html.

O-6. 2000 US STANDARD POPULATION

Starting with the November 2004 SEER submission of data (diagnoses through 2002), the SEER Program age-adjusts using the 2000 US standard population based on single years of age from the Census P25-1130 series estimates of the 2000 US population (Day, 1996). For the *CSR*, 19 age groupings were used for age-adjustment: <1, 1–4, 5–9, … , 80–84, 85+.

O-7. NUMBERS OF ESTIMATED CANCERS AND DEATHS IN 2008

The American Cancer Society (**ACS**) projects of the numbers of cancer cases and cancer deaths in the US in 2008 (American Cancer Society, 2008; Jemal et al., 2008). The ACS projects incidence in 2008 based on incidence rates for 1995-2004 from 41 states, representing about 85% of the US population. These high quality incidence data were submitted to the North American Association of Central Cancer Registries (NAACCR) by 41 states belonging to the SEER Program and/or the National Program of Cancer Registries (NPCR).

O-8. 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 death rates for the most recent 5-year period, trends (percent change (PC) and annual percent change (APC)) from 1975 to the most recent year, median age at diagnosis, median age at death, and survival rates. The information is provided by race (all races, white, and black) and by sex.

O-9. LONG-TERM TRENDS, 1950-2005

Trends in cancer mortality from 1950 to 2005 are summarized by age both for all cancers combined and for lung cancer (Table I-2). These cancer mortality trends are based on the mortality experience in the entire US. Summaries of long-term trends back to 1950 in cancer survival are also shown for whites.

Use caution when interpreting these statistics. Evaluating trends over a long period of time may

hide recent changes in the trends.

O-10. YEARS OF LIFE LOST DUE TO PREMATURE DEATH FROM VARIOUS CAUSES

Death rates alone give an incomplete picture of the burden that 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 a given age and sex. The life table permits a determination of the number of additional years an average person of that age, race, 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.

O-11. CANCER PREVALENCE

Methods: In this report prevalence is calculated at 1/1/2005. **Limited-duration prevalence** is calculated using the counting method implemented in the SEER*Stat software. This method calculates the number or proportion of people alive at the prevalence date who had a diagnosis of the disease within the past *x* years (e.g., *x* = 5, 10, 20, or the full history of the registry). Year of diagnosis is stratified into 5-year groups from the prevalence date, with the least recent interval being of varying length (4-8 years), depending on the length of years used to calculate prevalence. Race is stratified into white, black, other (American Indian/Alaska Native, Asian/Pacific Islander), and unknown/other-unspecified. When we use the SEER 11 registries, the same stratification as before is used, with American Indian/Alaska Native separated from Asian/Pacific Islander. Prevalence calculations for Hispanics use race stratified into: white, non-white, and unknown.

This method includes a correction for people lost to follow-up. For each individual lost to follow-up, a probability of being alive at the prevalence date is estimated from an appropriate survival function stratified by age at diagnosis (0–59, 60–69, 70+), sex, cancer site, year of diagnosis, and race, conditional on being alive at the time of loss to follow-up.

Because SEER has available information for the various racial/ethnic groups for different numbers of years, different years and registries were used to estimate prevalence. Prevalence estimates for all races combined, for whites, and for blacks use cases from 1975 through 2005 from the SEER 9 registries; prevalence estimates for Asian Pacific Islanders and Hispanics use

cases diagnosed from 1990 through 2005 from the SEER 11 areas and rural Georgia.

Different methods can be used to determine which tumors are to be included for people diagnosed with multiple tumors. Unless otherwise specified, prevalence calculations included only the *first malignant tumor per person*; that is, in situ cancers and second-or-later primary cancers were not included. Thus, if a woman had a melanoma prior to a breast cancer diagnosis, her melanoma would contribute to the prevalence of melanoma and to the prevalence of all sites, but the breast cancer would not contribute to the prevalence of breast cancer. Counting only one cancer per individual avoids some ambiguity in prevalence counts, and allows the counts for individual sites to sum to the all sites total. Prevalence using different selection criteria is compared in a table in the overview chapter. For more information on tumor selection criteria refer to http://srab.cancer.gov/prevalence/methods.html.

Complete prevalence is an estimate of the number of persons (or the proportion of population) alive on a specified date who had been diagnosed with the given cancer, no matter how long ago that diagnosis was. It was estimated for all races, whites, and blacks by applying the *completeness index method* (Capocaccia & De Angelis, 1997; Merrill et al., 2000; Mariotto et al., 2002) to limited-duration prevalence. The completeness index method is implemented in the COMPREV software (http://srab.cancer.gov/comprev/). Validation of the completeness index for all races and for whites was made by using data from the Connecticut Tumor Registry (CTR) beginning with 1940; for blacks, SEER 9 data beginning with 1975 were used. Identification of blacks is not possible in the CTR data prior to 1970. To validate the completeness index for blacks, we have compared the performance of the method to obtain 24-year prevalence from 10-year limited-duration prevalence. For all races combined and for whites, in cases where the validation indicated some lack of fit of the model, an approximation to the completeness index was derived from the CTR data. If there was a lack of fit for blacks, no estimate of complete prevalence was reported. Complete prevalence for Asian/Pacific Islanders and Hispanics is not available at this time. Complete prevalence by age for all races combined was validated by comparing estimated 10-year complete prevalence with observed prevalence from the CTR data. Prevalence by age is reported for the sites that validated well.

The US cancer prevalence counts at 1/1/2005 *were estimated* by multiplying the SEER age- and race-specific prevalence proportions by the corresponding US population estimates based on the average of 2004 and 2005 population estimates from the US Bureau of the Census. US cancer prevalence counts for all races were estimated by summing the US estimated counts for whites/unknown, blacks, and other races. For Hispanics, the estimates for Hispanics of white or unknown race and for Hispanics of other races were summed.

Limited-duration prevalence proportions by age at prevalence are not shown for childhood cancers (diagnosis before age 20) since many of these estimates are not informative. (For example, the number of people diagnosed with childhood cancers in the last 25 years and who are currently age 50-59 is zero by definition.) While it is of interest to estimate the total number of Americans currently alive who were diagnosed with a childhood cancer, the limitations of the

duration of the SEER cancer registries requires that this be estimated using statistical modeling. (This work is in progress.)

For more details on available prevalence estimates, see http://srab.cancer.gov/prevalence/.

Results and Table Description: The total number of persons alive on January 1, 2005, in the US who had had a diagnosis of invasive cancer is now estimated to be **11,098,450**. Compared with last year's 2004 prevalence estimate of **10,762,214** persons, this year's 2005 estimate represents an increase of **336,236** cases. This increase is due to increases in incidence, improvements in survival, and the increase and aging of the US population. The overview chapter contains two prevalence tables. The first table reports US complete prevalence counts by age at prevalence and sex for some cancer sites. The second table reports US prevalence counts for people diagnosed in the 5 years and 30 years prior to the prevalence date using different tumor inclusion criteria. Each site-specific chapter contains a prevalence table that reports limited-duration US prevalence counts by time since diagnosis for different racial/ethnic groups. US complete prevalence estimates are also reported when available. The second part of the table displays the percent of the population in the SEER 11 areas diagnosed in the previous 15 years with the specific cancer by 10-year age groups for the different racial/ethnic groups.

O-12. PROBABILITY OF BEING DIAGNOSED WITH OR DYING FROM CANCER

Each site-specific section 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 invasive cancer within the next 10, 20, or 30 years, or within their remaining lifetime. Lifetime risks of being diagnosed with invasive 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 2003-2005 incidence rates from the SEER 17 areas and death 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 that 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., between age 50 and age 60) 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. The methodology is described in detail in Fay (2003, 2004). To improve the precision of the calculations, rates were calculated beyond the usual last open ended age interval (i.e. 85+) for the age groups 85-89, 90-94, and 95+.

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. The estimates of developing and dying from cancer are implemented in DevCan (Probablity of DEVeloping or dying from CANcer software). More details on the software, various databases, and the methodology can be found at *http://srab.cancer.gov/devcan/*.

O-13. U.S. CANCER DEATH RATES BY STATE

Each cancer-site-specific section presents the death rate for the given cancer for each state and the District of Columbia, specifying the five highest and the five lowest death rates by state 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 2000 US standard million population. (In some previous editions of the CSR, the 1970 US standard million population was used; *death rates standardized to the 2000 US standard million population cannot be compared to death rates standardized to the 1970 US standard million population.*)

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

$$PD = [(\text{State Rate} - \text{Total US Rate})/\text{Total US Rate}] * 100$$

The **standard error** for each age-adjusted state death 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

$$SE_d = \text{Square Root of } [SE_S^2 + SE_U^2 - 2 * Cov_{S,U}]$$

where SE_S and SE_U are the standard errors of a state rate and of the total US rate, respectively, and $Cov_{S,U}$ is the covariance between the two rates. The variance of each rate (i.e., the square of the standard error) and the covariance between the two rates are 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, the statistical test may not be substantially affected if the state represents a small proportion of the total US. There is also an adjustment for multiple comparisons; see below under *Statistical Significance*.

O-14. JOINPOINT REGRESSION ANALYSIS OF CANCER TRENDS

An advance in the presentation of cancer trends is the use of joinpoint models (Kim et al., 2000). In some past issues of the *Cancer Statistics Review*, certain time intervals (e.g., 1973–1996) were specified and the annual percent changes (APC) 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, a statistical algorithm finds the optimal 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. A **joinpoint regression model** describes the trends by a sequence of connected segments where each segment is connected by a straight line on a log scale. Adjacent segments are connected at a joinpoint. The segments are connected because we assume that rates generally change smoothly, rather than "jump" abruptly. The rates are assumed to grow or decay exponentially, i.e., to change by a constant percentage each year. Thus the slope in each segment can be associated with a fixed annual percent change (**APC**).

Joinpoint analysis first assumes no joinpoints are needed to describe the data accurately, i.e., the trend over the entire interval 1975-2005 does not change. Joinpoints are added in turn if

they are statistically significant. Thus, in the final model, each joinpoint represents a significant change in trend. Computational considerations currently limit the maximum possible number of segments to be no larger than four, with three joinpoints. Smoother polynomial models may provide a good fit overall, but are less sensitive to what is occurring at the ends of the data.

In running the Joinpoint program, we set the program parameters as follows: maximum number of joinpoints 3, minimum interval between joinpoints 2 years, minimum interval between a joinpoint and an endpoint 2 years, joinpoints occurring only at exact years. These restrictions provide some added stability to the resultant models. Different values for these parameters may yield a different joinpoint model. Since the test statistic to determine if additional joinpoints are necessary cannot be compared against any known standard distribution to determine significance, (e.g., the normal, t, or f) a permutation test is used which simulates the distribution of the test statistic under the null hypothesis. Thus an element of randomness is introduced by the random number stream used. However, for greater consistency in the p-values obtained if one were to change the random seed for each run, we run the program for 4499 permutations.

Average Annual Percent Change (AAPC) is a summary measure of a trend over a pre-specified fixed interval based on an underlying joinpoint model. It allows us to use a single number to describe the average trend over a period of multiple years. It can be estimated even if the joinpoint model indicates that there were changes in trends during those years, since it is estimated as a weighted average of the joinpoint APCs, with the weights equal to the lengths of each segment over the pre-specified fixed interval. In this report, we have included AAPCs as an addendum to the underlying joinpoint trends, and as a summary measure to compare fixed interval trends by race/ethnicity. For more information on how the AAPC is calculated and the advantages of reporting an AAPC over APCs, see http://srab.cancer.gov/joinpoint/aapc.html.

A Windows-based program, *Joinpoint*, is freely available at http://srab.cancer.gov/joinpoint/; it accepts data from the *SEER*Stat* program, as well as user defined data. Further details on joinpoint regression may be found at the web site.

O-15. REPORTING DELAY

Timely and accurate calculation of cancer incidence rates is hampered by **reporting delay**, the time lapse before a diagnosed cancer case is reported to the NCI or the delay in receiving updated information for an existing case. Currently, the NCI allows a standard delay of 22 months between the end of the diagnosis year and the time the cancers are reported to the NCI in November, almost two years later. The data are released to the public in the spring of the following year. For example, cases diagnosed in 2005 were first reported to the NCI in November 2007 and released to the public in April 2008. However, in each subsequent release of the SEER data, *records from all prior diagnosis years* (e.g., diagnosis years 2004 and earlier in the 2007 submission to the NCI) *are updated* as either new cases are found or new information is received about previously submitted cases.

The submissions for the most recent diagnosis year are, in general, about two percent below the total number of cancers that will eventually be submitted for that year, although this varies by cancer site and other factors. The idea behind modeling reporting delay is *to adjust the recent rates to anticipate future corrections (additions, changes, and deletions) to the data*. These adjusted rates and the associated delay model are valuable in more precisely determining current cancer trends, as well as in monitoring the timeliness of data collection—an important aspect of quality control (Clegg et al., 2002). Reporting delay models have been previously used in the reporting of AIDS cases (Brookmeyer & Damiano, 1989; Pagano et al., 1994; Harris, 1990).

In this report, we show SEER age-adjusted incidence rates and trends, along with their calculated delay adjustments for all cancers combined (malignant only except for urinary bladder), for female breast in situ, for urinary bladder (in situ and malignant), and for 22 malignant cancer sites: melanoma (for all races combined and whites only), lung/bronchus, colon/rectum, prostate, female breast, liver and intrahepatic bile duct, pancreas, cervix uteri, corpus and uterus, ovary, testis, kidney and renal pelvis, brain and other nervous system, Hodgkin lymphoma, non-Hodgkin lymphoma, all leukemias, esophagus, larynx, myeloma, oral cavity and pharynx, thyroid, and stomach.

Cancer data from diagnosis years of 1981 to 2004 were used to model reporting delay distribution. A delay distribution models the probability of a cancer being reported after a delay of d years (d = 2, 3, ..., 24). The number of cancers reported at each delay year is assumed to follow a Poisson distribution. Cases are removed as corrections to the data are made, and the probability of removing cases is modeled as a binomial distribution. To reduce the number of parameters that have to be estimated and to achieve stability in the tails of the delay distributions, an assumption is made that all cancer cases will be reported within 25 years of diagnosis.

The delay distributions were modeled as a function of covariates using a discrete-time proportional hazards model. For the models presented here the following potential covariates are included: age at diagnosis, sex, diagnosis year, delay time, and race. Age at diagnosis was modeled as a 3-category variable with levels 0–49, 50–64, and 65+. Diagnosis year was modeled either as a continuous covariate or as categorized variables: 1981–1985, 1986–1990, and 1991–2004. Delay time d was modeled as a categorical variable in one of three ways:

(1) $d > 2$ or $d > 3$,
(2) $d > 2$, $d > 3$, $d > 4$, or $d > 5$, and
(3) $d > 2$, $d > 3$, ... , or $d > 10$.

For each cancer site, delay distributions were calculated for all races combined, for whites only, and for blacks only. When modeling delay distributions for all races combined, if a patient's race value changed between two submission years the change of value does not contribute to

the delay distribution. For melanoma, only all races combined and whites were analyzed because melanoma is rare for blacks.

Maximum likelihood estimates of delay probabilities were obtained using the Newton-Raphson algorithm. Details of the estimation can be found in Midthune et al. (2005). For each of the cancer sites, up to 72 models of pre-determined combinations of covariates were considered. We evaluated these models by fitting the models using data of diagnosis years between 1981 and 2004 and then predicting the cancer counts for 2005. For each cancer site, the model that minimized the sum of squared prediction errors was chosen as the default model. An algorithm was then used to compare the default model with competing models in order to select a model that best balanced prediction and simplicity. The chosen model was then refitted using all data (1981–2005 diagnosis years) to estimate delay distributions and calculate delay-adjusted estimates of the cancer rates.

Age-adjusted (using the 2000 US standard population) cancer incidence rates were then calculated with and without adjusting for reporting delay. Joinpoint linear regression (Harris, 1990) was used to obtain the annual percentage changes for the 1975–2005 incidence rates for the data series with and without delay adjustment. Incidence rates for diagnosis years prior to 1981 are not adjusted. Joinpoints were constrained to be at least two years away from both the beginning and the end of the data series and at least two years apart. Joinpoint regressions were fitted using the weighted-least-squares (weighted by appropriate variances of age-adjusted incidence rates) option in the *Joinpoint* regression software.

Results show that adjusting for delay tends to raise cancer incidence rates in more recent reporting years. While this adjustment increases the rate of change over the most recent diagnosis years, it probably will only rarely cause the detection of a new joinpoint, although this is possible. See Clegg et al. (2002) for details on the impact of reporting-delay adjustment to SEER cancer incidence rates.

For estimates of delay-adjusted rates, delay-adjustment factors, description of the covariates included in each cancer site model, and other details of delay adjustment, see *http://srab.cancer.gov/delay/*. The estimates of the delay-adjusted rates are in the Cancer Query Systems (*http://seer.cancer.gov/canques/*).

O-16. STATISTICAL SIGNIFICANCE

Errors may be made in the estimation 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 (usually 0.05), 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 α = 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 $P(\text{at least 1 success}) < p_1 + ... + p_n$) (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 $\alpha \approx 0.0002$. 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.

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 death 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 (58% of the total population in 2005) than any state. In addition, their higher mortality rates for several types of cancer elevate the overall rate for the District.

O-17. 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. For some sites, the historic stage is not shown because of inconsistencies in its definition over time or because stage isn't appropriate such as leukemias which are all considered to be distant/disseminated at diagnosis.

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. 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, 1990, and 2000 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 *CSR*s prior to 1973–1996.

Comparison with other databases: The SEER data are obtained from population-based cancer registries covering about 26 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 or standard million population should be used for the age-adjustment of each group being compared. Examples of other databases are USCS (US Cancer Statistics Working Group, 2005) and CINA+ Online (*http://www.naaccr.org/cinap/*).

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 US population; in some previous SEER reports, the 1970 standard million population was used. Therefore, *rates in this report can not be compared to rates and trends in those 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 in a given calendar year may change from that which has 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 revised and updated by the Bureau of the Census (**BOC**). Such changes may result in some differences between incidence and mortality rates for a given calendar period as published in different reports. See our website for the most current information about the population estimates (*http://seer.cancer.gov/popdata/*).

O-18. 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):

$$SE(CR_t) = CR_t * \text{square root of } [q_1/(e_1-d_1) + q_2/(e_2-d_2) + \ldots + q_t/(e_t-d_t)]$$

where CR_t is the t-year relative survival rate, and for $i = 1, \ldots, t$,
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 SE of a crude incidence or mortality rate (Keyfitz, 1966), that is, the SE can be approximated by the rate divided by the square root of the number of cancer cases (or the number of deaths).

Appendix tables 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 replace the number of cancer cases or deaths with the number of cancer cases or deaths divided by 5.

O-19. 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 persons at risk. That is,

Incidence rate = (New cancers / Population) * 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 primary site reported is the site of origin and not the metastatic site. In general, the incidence rate would not include recurrences. *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 in this report are *age-adjusted* (see below) to the 2000 US standard population (or, where appropriate, to the world standard million population). (In some previous editions of the *CSR*, the 1970 US standard million population was used; therefore, *incidence rates in this edition cannot be compared to rates published in those editions.*) Incidence rates are for *invasive cancer only*, unless otherwise specified. (Exceptions are the incidence rate for cancer of the urinary bladder (where both in situ and invasive cancers are counted) and breast cancer in situ, which is shown separately.)

Death rate: The cancer death (or 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 persons. That is,

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

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

Age distribution: A table showing a partition of the entire lifespan into disjoint age intervals, along with the proportion of the population in each interval.

Median age: The age at which half of a population is younger and half is older.

Standard 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 standard population. Table A-7 shows the US 2000 standard population and the world standard million population. (Some World Health Organization mortality publications use a different world standard million population.)

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 counts 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 based on the same standard million population. For this report, the 2000 US standard million population (or, where appropriate, the

world standard million population) is used in computing age-adjusted rates, unless otherwise noted.

Percent change: The percent change (**PC**) in a statistic over a given time interval is

Percent change = (Final value – Initial value) / Initial value * 100.

A positive PC corresponds to an increasing trend, a negative PC to a decreasing trend.

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

Because the methods used in their calculation are mathematically different, *the signs of the PC and the APC 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*, called 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.

Average Annual Percent Change: The average annual percent change (**AAPC**) is a summary measure of a trend over a pre-specified fixed interval based on an underlying joinpoint model. It allows us to use a single number to describe the average trend over a period of multiple years. It can be estimated even if the joinpoint model indicates that there were changes in trends during those years, since it is estimated as a weighted average of the joinpoint APCs, with the weights equal to the lengths of each segment over the pre-specified fixed interval.

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 cohort 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, race, 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 for a given year is simply the sum of this expectation over all those individuals who died of that cancer in that year.

Average years of life lost: The average years of life lost (**AYLL**) associated with a particular cancer for a given year 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 that 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 (incident) and pre-existing cases and is a function of past incidence, past survival, and the size and age structure of the population**.** *Limited-Duration Prevalence* represents the proportion of people alive on a certain day who had a diagnosis of the disease within the past *x* years (e.g. *x* = 5, 10, or 20 years). *Complete prevalence* is an estimate of the number of persons (or the proportion of the population) alive on a specified date who had been diagnosed with the given disease, no matter how long ago that diagnosis was. For more details on cancer prevalence definitions and methods, refer to *http://srab.cancer.gov/prevalence/*.

Stage of disease at diagnosis: Extent-of-disease information determines stage of disease at diagnosis. The **SEER historic 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).

O-20. REFERENCES

American Cancer Society. Cancer Facts and Figures 2008. Atlanta: American Cancer Society, 2008.

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.

Beahrs 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.

Brookmeyer R, Damiano A. Statistical methods for short-term projections of AIDS incidence. Statistics in Medicine 1989;8:23-34.

Byrne J, Kessler LG, Devesa SS. The prevalence of cancer among adults in the United States: 1987. Cancer 1992;68:2154-9.

Capocaccia R, De Angelis R. Estimating the completeness of prevalence based on cancer registry data. Statistics in Medicine 1997;16:425-40.

Clegg LX, Feuer EJ, Midthune D, Fay MP, Hankey BF. Impact of reporting delay and reporting error on cancer incidence rates and trends. J Natl Cancer Inst 2002;94:1537-1545.

Clegg LX, Mitchell GH, Feuer EJ. Estimating the variance of disease prevalence estimates from population-based registries. Technical report 2001. For a copy contact prevalence@ims.nci.nih.gov.

Day, Jennifer Cheeseman, *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050*, U.S. Bureau of the Census, Current Population Reports, P25-1130, U.S. Government Printing Office, Washington, DC, 1996. (*http://www.census.gov/prod/1/pop/p25-1130/p251130.pdf*)

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.

Fay MP. Estimating age conditional probability of developing disease from surveillance data. Popul Health Metr 2004 Jul 27;2(1):6. [http://www.pophealthmetrics.com/content/2/1/6]

Fay MP, Pfeiffer R, Cronin KA, Le C, Feuer EJ. Age-conditional probabilities of developing cancer. Stat Med 2003;22(11):1837-48.

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.

Gail MH, Kessler L, Midthune D, Scoppa S. Two approaches for estimating disease prevalence from population-based registries of incidence and total mortality. Biometrics 1999;55:1137-44.

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.

Harris JE. Reporting delays and the incidence of AIDS. Journal of the American Statistical Association 1990;85:915-924.

Ingram DD, Parker JD, Schenker N, Weed JA, Hamilton B, Arias E, Madans JH. United States Census 2000 population with bridged race categories. Vital Health Stat 2. 2003 Sep;(135):1-55.

Jemal A, Siegel R, Ward E, Hao Y, Xu J, Murray T, Thun MJ. Cancer statistics, 2008. CA Cancer J Clin 2008;58:71-96.

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. Mariotto A, Gigli A, Capocaccia R, Clegg L, Scoppa S, Ries LA, Tesauro GS, Rowland JS, Feuer EJ. Complete and limited duration prevalence estimates. *SEER Cancer Statistics Review, 1973-1999* 2002;19.

Merrill RM, Feuer EJ, Capocaccia R, Mariotto A. Cancer prevalence estimates based on tumor registry data in the SEER Program. Int J Epidemiol 2000;29:197-207.

Midthune DN, Fay MP, Clegg LX, Feuer EJ. Modeling reporting delays and reporting corrections in cancer registry data. J Am Stat Assoc 2005;100(469):61-70.

Pagano M, Tu XM, De Gruttola V, & MaWhinney S. Regression analysis of censored and truncated data: estimating reporting-delay distributions and AIDS incidence from surveillance data. Biometrics 1994;50:1203-1214.

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.

Robinson JG, West KK, Adlakha A. Coverage of the population in Census 2000: Results from demographic analysis. Population Res Policy Rev 2002;21:19-38.

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 Cancer Statistics Working Group. *United States Cancer Statistics: 1999-2002 Incidence and Mortality Web–based Report Version*. Atlanta: Department of Health and Human Services, Centers for Disease Control and Prevention, *and* National Cancer Institute; 2005. Available at: http://www.cdc.gov/cancer/npcr/uscs.

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, Mattheiem WH, Rozencweig M, editors. Breast Cancer: Trends in Research and Treatment. New York (NY): Raven Press; 1976. p. 287-299.