these life tables will be used as the default to estimate expected survival for that only include cases diagnosed after 1992 (for example SEER (2000+).

The state/race/SES life table were constructed using counts of deaths and populations by county, single year age at death (30 to 84 years), race/ethnicity, sex, and calendar year 1992-2013. We used mutually exclusive race/ethnicity groups: Non-Hispanic (NH) White, NH Black, NH AIAN, NH API, and Hispanics (hereafter we exclude the NH prefix when referencing race/ethnicity). Hispanic ethnicity includes all race categories. Because of misclassification errors of AIAN race in death certificates, we restricted the AIAN data to mortality rates from Purchased/Referred Care Delivery Areas (PRCDA) counties. We fit Poisson regression models to the log of mortality rates to estimate the life tables separately for men and women and each race/ethnicity. Age and calendar year were modeled as spline functions to capture non-linear effects. The models varied by geographic area (state, region, and national) and the inclusion or not of the SES index as a covariate depending on sufficient numbers of deaths and population counts for each race-ethnicity. For more details on the methods and data to estimate life tables a technical is available on request.

CAUSE-SPECIFIC SURVIVAL

Cause-specific survival is a net-survival measure representing survival of a specified cause of death in the (theoretical) absence of other causes of death. Estimates are calculated by specifying the cause of death. Individuals who die of causes other than the specified cause are censored. This requires a cause-of-death variable that accurately captures all causes related to the specific cause. Cancer registries use algorithms to process causes of death from death certificates in order to identify a single, disease-specific, underlying cause of death. In some cases, attribution of a single cause of death may be difficult and misattribution may occur. For example, a death may be attributed to the site of metastasis instead of the primary site (Percy et al., 1981).

To capture deaths related to the specific cancer but not coded as such, the SEER cause-specific death classification variable is defined by taking into account causes of deaths in conjunction with tumor sequence (i.e., only one tumor or the first of subsequent tumors), site of the original cancer diagnosis, and comorbidities (e.g., AIDS and/or site-related diseases). To learn more on this topic, please read the recent article published at the Journal of National Cancer Institute (Howlader et al., 2010) or visit: https://seer.cancer.gov/causespecific/.

CANCER PREVALENCE

METHODS: In this report prevalence is calculated at 1/1/2017. 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). With the release of the 1975-2017 Cancer Statistics Review, the calculation of all complete and limited-duration prevalence estimates were modified to use data from the SEER 13 areas not including the Alaska Natives Registry using cases diagnosed from 1992 through 2017.

The limited-duration prevalence 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. 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 13 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.

Different methods can be used to determine which tumors are to be included for people diagnosed with multiple tumors. In previous reports published in 2016 and before a different method was used: 1st invasive tumor ever of a person. This method only includes people for their first tumor ever. Unless otherwise specified, prevalence calculations include the first invasive tumor per cancer site for the total prevalence duration. In this method, the first invasive tumor per cancer site diagnosed during the total prevalence duration can contribute to cancer prevalence statistics. For example, if a woman had a melanoma diagnosed in 1992, a breast cancer diagnosed in 2000 and a second breast cancer diagnosed in 2005, her melanoma will contribute to the prevalence of melanoma and to the prevalence of all sites, and the first breast cancer will contribute to the prevalence of breast cancer. However, if we are calculating 16-years prevalence including individual's first cancer per site between 2000-2015 the melanoma diagnosed in 1992 would not contribute to 16-year melanoma prevalence and the 2000 breast cancer will contribute to the all sites and breast prevalence. Because prevalence counts people and not tumors, the women is included once in the breast cancer prevalence for her first breast cancer. In the 1st invasive tumor ever the woman's melanoma cancer 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. For more information on tumor selection criteria refer to http://surveillance.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, which can be found at https://surveillance.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 1/1/2017 *were estimated* by multiplying the SEER age- and race-specific prevalence proportions by the corresponding US population estimates based on the average of 2016 and 2017 population estimates from the US Census Bureau. 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.
Complete prevalence estimates of the number of individuals in the US diagnosed with cancer as children (ages 0-19), including those surviving for more than 24 years, is calculated using a statistical method that estimates the number of childhood survivors diagnosed before 1992 (Simonetti et al., 2008; Mariotto et al., 2009). Limited-duration prevalence proportions by age at prevalence are not shown for childhood cancers (age at diagnosis 0-19) 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. For more details on available prevalence estimates, see https://surveillance.cancer.gov/prevalence/.

PROBABILITY OF BEING DIAGNOSED WITH OR DYING FROM CANCER

LIFETIME AND INTERVAL RISKS OF BEING DIAGNOSED WITH CANCER: The probability of being diagnosed with cancer is computed by applying cross-sectional age-specific 2015-2017 incidence rates from the SEER 21 areas and death rates from those same areas 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 et al., 2003) and (Fay, 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.

DETAILED METHODOLOGY AND SOFTWARE: 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 https://surveillance.cancer.gov/devcan/.

US 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 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*.

The states are ranked according to the death rate, with 1 indicating the highest and 51 the lowest rate in the US. 95% confidence intervals for the rank are shown in parentheses () after the rank. The confidence intervals of ranks of age-adjusted rates are calculated using a simulation-based method (Zhang, 2014) implemented in the CI*Rank tool https://surveillance.cancer.gov/cirank/.

JOINPOINT REGRESSION ANALYSIS OF CANCER TRENDS

Joinpoint regression is a useful way to characterize trends in cancer rates and other heath indices (Kim et al., 2000). It characterizes segments using connected linear segments on a log scale (i.e. constant annual percent changes (APC's) between changepoints. The locations of the changepoints are optimally determined using by the data using a statistical algorithm. To achieve greater descriptive accuracy, a statistical algorithm finds the optimal number and location of places where a trend changes. The point (in time) when 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 continuous, piecewise-exponential function.

Adjacent segments are connected at a joinpoint. The segments are connected because we assume that rates generally change smoothly, rather than "jump" abruptly. In each segment, the rates are assumed to grow or decay exponentially ($y = e^{mx+b}$), i.e., to change by a constant percentage each year. Thus the "slope" *m* in each segment can be associated with a fixed annual percent change (**APC**) by $APC = 100(e^m - 1)$.

Joinpoint analysis first assumes no joinpoints are needed to describe the data accurately, i.e., the trend over the entire interval 1975-2014 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. 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:

(1) Joinpoints occur only at exact years; the joinpoint is not necessarily the same as the data point for that year;
(2) The minimum time interval between consecutive joinpoints is three years;
(3) The first joinpoint is not earlier than two years after the first year of data;
(4) The last joinpoint is not later than two years before the last year of data;
(5) The maximum number of joinpoints is five for 1975-2013 (SEER 9) data and three for 1992-2014 (SEER 13) data.

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.

A Windows-based program, *Joinpoint*, is freely available at https://surveillance.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 website. Starting with the 2012 edition of CSR, we have generated all our cancer trend statistics using a Linux-based *Joinpoint* program as opposed to the downloadable Windows-based program. As a result of using a different platform, in rare instances the results (e.g., # of joinpoints) may differ.

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 geometric 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 https://surveillance.cancer.gov/help/joinpoint/setting-parameters/method-and-parameters-tab/apc-aapc-tau-confidence-intervals/average-annual-percent-change-aapc.

JUMP MODEL/COMPARABILITY RATIO MODEL

The Jump Model / Comparability Ratio Model in the Joinpoint software provides a direct estimation of trend data (e.g. cancer rates) where there is a coding, which causes a "jump" in the rates, but is assumed not to affect the underlying trend. To account for ICD-9 to ICD-10 coding change, occurred in 1998, alternative trends estimated from Jump model and Comparability Ratio Model are obtained for Melanoma. Those trends and more information can be found in https://surveillance.cancer.gov/joinpoint/jump.html .

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, 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 2014 were first reported to the NCI in November 2016 and released to the public in April 2017. However, in each subsequent release of the SEER data, *records from all prior diagnosis years* (e.g., diagnosis years 2014 and earlier in the 2016 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. To adjust for this, statistical models have been developed to estimate "reporting delay-adjusted rates" for the SEER 9 since 2003 and SEER 13 registries since 2010 and the delay adjusted rates are reported.

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).

In addition to registries funded by NCI-SEER, registries for the remainder of the U.S. are funded by the Centers for Disease Control and Prevention National Program of Cancer Registries (CDC-NPCR). (Some registries are co-funded by both NCI and CDC). Annual cancer incidence

and survival data are reported by U.S. registries to NCI-SEER and CDC-NPCR, while registries throughout the US and Canada are report annually to the North American Association of Central Cancer Registries (NAACCR), a registry member organization. A coordinated effort by NCI, CDC and NAACCR has led to a unified approach to estimate and report delay adjusted rates.

Starting with data released in 2015, for the first time, delay adjustment factors is produced based on December 2014 data submitted to the NAACCR. The delay adjusted rates are then estimated from the delay adjustment factors by cancer site, registry, age group, gender, race, and year of diagnosis and linked to the appropriate cases (based on cancer site, registry, age group, gender, race, and year of diagnosis), to data submissions for each of the three partners in this joint effort (NCI-SEER, NAACCR, and CDC-NPCR). Starting from 2017 release, delay adjustment factors for Ethnicity (Hispanic and Non-Hispanic) and Race x Ethnicity combination are also estimated. This will allow all the partners and users of these data to produce delay adjusted rates. See Appendix for details.

In this report, we show SEER age-adjusted incidence rates and trends, along with their calculated delay adjustments for SEER 9 and SEER 13 areas. The adjusted rates, factors, and trends are available for all cancers combined (malignant only except for urinary bladder), for female breast in situ, for urinary bladder (in situ and malignant combined), 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 leukemia, esophagus, larynx, myeloma, oral cavity and pharynx, thyroid, and stomach.

For more information on cancer incidence rates adjusted for reporting delay, see https://surveillance.cancer.gov/delay/.

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 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 —51% of the total population in 2010 (US Census Bureau, 2013)—than any state. In addition, their higher mortality rates for several types of cancer elevate the overall rate for the District.

STANDARD ERRORS OF RATES

SURVIVAL RATES: In the tables presenting survival estimates, the magnitude of the standard error is given as a measure of the reliability of a given rate: the greater the standard error, the more uncertainty associated with the estimated 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 estimate could not be calculated, as is noted in the table footnotes.

The **standard error** (**SE**) of a relative survival estimate 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 estimate, 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.

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,

Death Rate = (Cancer Deaths / 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 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 population. The potential confounding effect of age is reduced when comparing age-adjusted rates based on the same standard population. For this report, the 2000 US standard 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 subinterval 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: The observed survival estimate 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: The relative survival estimate is calculated using a procedure (Ederer et al., 1961; Ederer and Heise, 1959) whereby the observed survival estimate is adjusted for expected mortality. The relative survival estimate 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 estimate 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 https://surveillance.cancer.gov/prevalence/.

STAGE OF DISEASE AT DIAGNOSIS: Extent-of-disease information determines stage of disease at diagnosis. The **SEER summary 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 (Greene et al, 2002; Edge et al., 2010).

SOFTWARE USED TO GENERATE THE SEER CANCER STATISTICS REVIEW

The SEER Cancer Statistics Review includes statistics generated by a variety of statistical software including:

- SEER*Stat, statistical software for the analysis of SEER and other cancer databases, was used to generate incidence, mortality, prevalence, and survival statistics presented in the CSR.
- Analysis generated by the Joinpoint Regression Program are presented to better describe trends that are not constant over time.
- The DevCan system generated the probability of developing cancer from twelve SEER areas and the probability of dying from cancer from the total United States.
- The ComPrev software was used to calculate complete prevalence estimates.

Additional statistics can be obtained via SEER's Cancer Query Systems. These data retrieval applications provide access to pre-calculated cancer statistics stored in online databases.

REFERENCES

American Cancer Society. *Cancer Facts & Figures 2012*. Atlanta: American Cancer Society; 2012.

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.

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. *Stat Med* 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. *Stat Med* 1997;16:425-40.

Cho H, Howlader N, Mariotto AB, Cronin KA. **Estimating relative survival for cancer patients from the SEER Program using expected rates based on Ederer I versus Ederer II method**. Surveillance Research Program, National Cancer Institute; 2012. Technical Report #2012-01.

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 L, Gail M, Feuer EJ. Estimating the variance of disease prevalence estimates from population-based registries. *Biometrics* 2002;58(3):684-8.

Day JC. *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1995 to 2050*, US Census Bureau, Current Population Reports, P25-1130, US Government Printing Office, Washington, DC, 1996. Available from: https://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.

Ederer F, Heise H. Instructions to IBM 650 Programmers in Processing Survival Computations, Technical, End Results Evaluation Section, National Cancer Institute, 1959.

Edge SB, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A III. *AJCC Cancer Staging Manual*, 7th ed. New York (NY): Springer; 2010.

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. Available from: https://pophealthmetrics.biomedcentral.com/articles/10.1186/1478-7954-2-6

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. 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.

Greene FL, Page DL, Fleming ID, Fritz AG, Balch CM, Haller DG, Morrow M, editors. *AJCC Cancer Staging Manual*, 6th ed. New York (NY): Springer; 2002.

Hahn RA, Mulinare J, Teutsch SM. Inconsistencies in coding of race and ethnicity between birth and death in US infants. *JAMA* 1992;267:259-263.

Hakulinen T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* 1982;38:933-942.

Harris JE. Reporting delays and the incidence of AIDS. *J Am Stat Assoc* 1990;85:915-924.

Howlader N, Ries LAG, Mariotto AB, Reichman ME, Ruhl J, Cronin KA. Improved estimates of cancer-specific survival rates from population-based data. *J Natl Cancer Inst* 2010;102:1-15.

Howlader N, Ries LAG, Stinchcomb DG, Edwards BK. The impact of underreported Veterans Affairs data on national cancer statistics: analysis using population-based SEER registries. *J Natl Cancer Inst* 2009;101(7):533-536.

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.

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: PWS-Kent, 1988.

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, Stanek E, Gloeckler L. Accuracy of cancer death certificates and its effect on cancer mortality statistics. . *Am J Public Health* 1981;71: 3242-3250.

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.

Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. *CA Cancer J Clin* 2012 Jan-Feb;62(1):10-29.

Simonetti A, Gigli A, Capocaccia R, Mariotto A. Estimating complete prevalence of cancers diagnosed in childhood. *Stat Med* 2008 Mar 30;27(7):990-1007.

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 from: https://www.cdc.gov/cancer/npcr/uscs/index.htm

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

US Census Bureau: State and County QuickFacts. Data derived from Population Estimates, Census of Population and Housing, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits, Consolidated Federal Funds Report. Last Revised: Thursday, 04-Nov-2010 12:46:18 EDT.

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.

Zhang, S, et al. Confidence intervals for ranks of age-adjusted rates across states or counties. *Statistics in Medicine*. 33(11): 1853-66.

Zou J, Huang L, Midthune D, Horner MJ, Krapcho M, Feuer EJ. Effect of reporting year on delay modeling. Statistical Research and Applications Branch, National Cancer Institute; 2009. Technical Report #2009-01. Available from: https://surveillance.cancer.gov/reports/.