BACKGROUND AND DATA SOURCES

The primary measures associated with assessing the impact of cancer in the general population are the number of new cases per year per 100,000 persons (incidence rate), the number of deaths per 100,000 persons (mortality rate), and a determination of the proportion of patients alive at some point subsequent to the diagnosis of their cancer (survival rate). All three measures are included in this report using data from the Surveillance, Epidemiology, and End Results (SEER) Program based within the Cancer Surveillance Research Program at the National Cancer Institute (NCI) and cancer mortality data provided by the National Center for Health Statistics (NCHS) for the entire United States (U.S.). All incidence and mortality rates in this report are age-adjusted to the 1970 United States standard million (see Appendix) unless otherwise specified. Age-adjustment minimizes the effect of a difference in age distributions when comparing rates.

The SEER Cancer Statistics Review (CSR), containing the most recent cancer incidence, mortality and survival statistics, is made available by the Cancer Statistics Branch of the NCI annually. Since 1996, the CSR has been available electronically on the SEER Home Page, http://www-seer.ims.nci.nih.gov, under "Publications." The Web page allows for the more timely distribution of the CSR. This CSR includes incidence, mortality, and survival data from 1973 through 1996, the most recent year for which data are available. Incidence data for 1996 appear to be 96 to 98 percent complete. Therefore, caution must be exercised when comparing rates for 1996 with those for previous years. Data are presented for a wide spectrum of cancers. The scope and purpose of this review are consistent with a report to the Senate Appropriations Committee (Breslow, 1988) which recommended that a broad profile of cancer be presented to the American public on a routine basis. Additional SEER data can be obtained via CANQUES, an interactive system on the SEER Web page under Scientific Systems, which allows the user to access over 10 million pre-calculated cancer statistics. The SEER public-use file with SEER*Stat is also available for ordering through Scientific Systems' part of the SEER Web page. SEER*Stat provides an easy to use PC desktop system for the production of a myrad of cancer statistics, such as incidence rates and survival rates by various demographic and tumor variables. The SEER public-use data file contains information on over two million tumors and contains no personal identifiers.

Incidence and survival data: The National Cancer Act of 1971 mandated the collection, analysis and dissemination of data useful in the prevention, diagnosis and treatment of cancer. This mandate led to the establishment of the SEER Program. A continuing project of the NCI, the SEER Program collects cancer data on a routine basis from designated population-based cancer registries in various areas of the country. Trends in cancer incidence, mortality and patient survival in the United States are derived from this database.

A sequel to two earlier NCI programs--the End Results Program and the Third National Cancer Survey--the SEER Program was initiated in several geographic areas of the United States and its territories, with case ascertainment beginning with January 1, 1973 diagnoses. 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.

In 1974-1975, the program was expanded to include the metropolitan area of New Orleans, Louisiana, the thirteen-county Seattle-Puget Sound area in the State of Washington and the metropolitan area of Atlanta, Georgia. New Orleans participated in the program only through the 1977 data collection year. In 1978 ten predominantly black rural counties in Georgia were added. American Indian residents of Arizona were added in 1980. In 1983, four counties in New Jersey were added with coverage retrospective to 1979. New Jersey and Puerto Rico participated in the program only until the end of the 1989 reporting year. Two areas of California, Los Angeles County and the San Jose-Monterey area (Monterey, San Benito, Santa Clara and Santa Cruz Counties) began reporting with 1992 diagnoses. Both population-based cancer registries began data collection earlier than 1992 and they have provided earlier data from 1988 through 1991 for inclusion in the CSR. The incidence trends and survival data for this report are from five States: Connecticut, Hawaii, Iowa, New Mexico, and Utah and four metropolitan areas: Detroit SMSA, Atlanta SMSA, San Francisco-Oakland SMSA, and Seattle-Puget Sound (Fig. I-1). Incidence rates by SEER area including Los Angeles and San Jose-Monterey are shown for the most recent 5-year period along with

area-specific mortality in each section.

Data from the nine or eleven SEER geographic areas used in this report represent an estimated 9.5 or 13.9 percent of the United States population, respectively. By the end of 1996, the database contained information on over 2 million cases diagnosed since 1973; currently over 125,000 new cases are accessioned yearly.

Areas were selected primarily 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 reasonably representative subsets of the United States population.

The goals of the SEER Program are:

1. Assembling and reporting, on a periodic basis, estimates of cancer incidence and mortality in the United States.

2. Monitoring annual cancer incidence trends to identify unusual changes in specific forms of cancer occurring in population subgroups defined by geographic and demographic characteristics.

3. Providing continuing information on changes over time in the extent of disease at diagnosis, trends in therapy, and associated changes in patient survival.

4. Promoting 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.

The SEER Program is conducted under contract with nonprofit, medically oriented organizations having statutory responsibility for registering diagnoses of cancer among residents of their respective geographic coverage areas. Each contractor:

1. maintains a cancer information reporting system;
2. abstracts records for resident cancer patients seen in every hospital in and outside the coverage area;
3. abstracts all death certificates on which cancer is listed as a cause of death for residents dying in and outside the coverage area;
4. searches records of private laboratories, radiotherapy units, nursing homes and other health services units which provide diagnostic service to ensure complete ascertainment of cases; registers all in situ and malignant neoplasms with the exception of certain histologies for cancer of the skin and in situ of the cervix uteri since 1996;
5. 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;
6. provides active follow-up on all living patients except for those with in situ cancer of the cervix uteri;
7. maintains confidentiality of patient records;
8. submits data electronically to NCI twice each year containing data on all reportable diagnoses of cancer which were made in residents of the coverage area.

In situ cancers of the cervix uteri are not reportable to SEER beginning with 1996 diagnoses. Since 1992, the SEER program has coded site and histology by the International Classification of Diseases for Oncology, second edition (ICD-O-2) (Percy, Van Holten et al, 1990). All cases before 1992 were machine converted to ICD-O-2. The primary site groupings used for incidence are found in the Appendix. Follow-up rates are also in the Appendix.

Mortality data: A public use file containing information on all deaths occurring in the United States by calendar year is obtained annually from the National Center for Health Statistics (NCHS). Information on each death includes age at death, sex, geographic area of residence, underlying and contributing causes of death. Only the underlying cause of death was used in the calculation of mortality rates. Numbers and the numerators for mortality rates for the SEER geographic areas, for each state and for the total U.S. are obtained from these tapes. A list of the mortality site groupings used in this publication is in the

Appendix.

Number of estimated cancers and deaths in 1999: Projections of the number of cancer cases and number of cancer deaths in the United States for 1999 have been obtained from the American Cancer Society (ACS). The ACS projected incidence to 1999 based on incidence rates from SEER for 1979-95 and applied by the ACS to the 1999 estimated total U.S. population (American Cancer Society, 1999).

Population data: Population estimates are obtained each year from the U.S. Bureau of the Census. Currently, revised estimates of the populations of U.S. counties were obtained by five-year age group (0-4, 5-9,..., 85 and over), sex, and race (including white and black) for July estimates for each year 1990- 1997. SEER makes county estimates for each state available on the SEER Home Page
(http://www-seer.ims.nci.nih.gov)
for race (whites, blacks, non-white), 5-year age group, sex, and year of diagnosis (each year 1973 to 1996). Additional racial/ethnic (Asian/Pacific Islander, American Indian, and Hispanic) populations for 1990-1997 by county can be obtained from the U.S. Census Bureau Web page, http://www.census.gov/population/www/estimates/countypop.html.

U.S. Bureau of the Census (BOC) population estimates for Hawaii were altered according to independent estimates developed from sample survey data collected by the Health Surveillance Program (HSP) of the Hawaii Department of Health. For Hawaii, the all races and black populations are the same as those sent by the BOC. Proportions of the population by different racial groups from the HSP were used to generate estimates for whites, etc. Since the HSP survey was for all of Hawaii and not by county, population estimates were not broken down by county. The white population estimates for Hawaii provided by the BOC are generally larger than those generated by the HSP. Since whites in Hawaii account for less than two percent of the total white population represented by the SEER reporting areas, white incidence rates for the entire SEER Program are not noticeably affected. Procedures for calculating rates by race for Hawaii are currently under review.

DEFINITIONS

Several technical terms are used in presenting the data in this report. The following definitions are presented here in an attempt to clarify their use to 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, expressed as the number of cancers per 100,000 people. It should be noted that the numerator of the rate can include multiple primary cancers occurring in one individual. This rate can be computed for each type of cancer as well as for all cancers combined. Except for five-year age-specific rates, all incidence rates are age-adjusted to the 1970 U.S. standard population or to the world standard (see below). Rates are for invasive cancer only, unless otherwise specified. For example, the rates for cancer of the urinary bladder are comprised of in situ and invasive cancer.

Mortality rate: The cancer mortality rate is the number of deaths with cancer given as the underlying cause of death occurring in a specified population during a year, expressed as the number of deaths due to cancer per 100,000 people. This rate can be computed for each type of cancer as well as for all cancers combined. Except for age-specific rates, all mortality rates are age-adjusted to the 1970 U.S. standard population or to the world standard (see below).

Age-adjusted rate: An age-adjusted rate is a weighted average of the age-specific cancer incidence (or mortality) rates, where the weights are the proportions of persons in the corresponding age groups of a standard population. The potential confounding effect of age is reduced when comparing age-adjusted rates computed using the same standard population. For this report, the 1970 United States standard million and world standard million populations are used as the standards in computing age-adjusted rates unless otherwise noted.

Percent Change: The percent change in rates over the entire time period covered by this report was obtained by calculating the average of the 1973 and 1974 rates and the average of the rates for the most recent two years, subtracting the former from the latter, dividing the difference by the former, and then

multiplying by 100 to convert the number to a percent. Percent changes are also provided for two five-year periods, 1975-79 and the most recent 5-year period.

Estimated Annual Percent Change: The Estimated Annual Percent Change (EAPC) was calculated by fitting a regression line to the natural logarithm of the rates (r) using calendar year as a regressor variable, i.e. $y = mx + b$ where $y = \text{Ln } r$ and x = calendar year. The EAPC = $100*(e^m - 1)$. Testing the hypothesis that the Annual Percent Change is equal to zero is equivalent to testing the hypothesis that the slope of the line in the above equation is equal to zero. The latter hypothesis is tested using the t distribution of m/SE_m with the number of degrees of freedom equal to the number of calendar years minus two. The standard error of m, i.e. SE_m, is obtained from the fit of the regression (Kleinbaum, 1988). This calculation assumes that the rates increased/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 equal to zero, the linear regression was not calculated. Because the methods used in their calculation are not directly related, it is possible that the signs of the PC and the EAPC may disagree, and this occurs in a few instances in the tables presented. The differences between incidence and mortality trends for the time period 1975-79 versus those for the most recent five-year period are tested for statistical significance using a t statistic with six degrees of freedom defined as the difference in the regression coefficients divided by the standard error of the difference (Kleinbaum, 1988).

Observed survival rate: The observed survival rate is obtained using standard life table procedures and represents the proportion of cancer patients surviving for a specified length of time after diagnosis.

Relative survival rate: The relative survival rate is calculated using a procedure described by Ederer, Axtell, and Cutler (1961) whereby the observed survival rate is adjusted for expected mortality. The relative survival rate represents the likelihood that a patient will not die from causes associated specifically with their cancer at 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) was calculated as follows. For each of the individuals who died of a particular cancer of interest, it was possible to obtain the number of additional years they were expected to survive, conditional on their survival to the age at which they died of their cancer. This conditional expectation was obtained from life tables for the United States population available from the National Center for Health Statistics. The PYLL in the general population associated with a particular cancer is simply the sum of this conditional expectation over all those individuals who died of that cancer.

Average Years of Life Lost: The Average Years of Life Lost (AYLL) associated with a particular cancer is the PYLL associated with that cancer in the general population divided by the number of deaths from that cancer in the general population.

Stage of Disease at Diagnosis: Localized - an invasive neoplasm confined entirely to the organ of origin. Regional - a neoplasm that has extended beyond the limits of the organ of origin directly into surrounding organs or tissues; into regional lymph nodes; or both direct extension and regional lymph node involvement. Distant - a neoplasm that has spread to parts of the body remote from the primary tumor either by direct extension or by discontinuous metastasis. Unstaged - information is not sufficient to assign a stage.

SUMMARY TABLES

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

LONG-TERM TRENDS, 1950-1996

Trends in cancer mortality from 1950 to 1996 are summarized by age for all cancers combined, excluding and including lung cancer (Table I-2). These mortality figures are based on the experience in the total United States.

Summaries of long-term trends in cancer incidence, mortality and survival are outlined in Table I-3. The table shows the estimated number of cancer cases and the reported number of cancer deaths for 1996; the next four columns show incidence and mortality changes from 1950 to 1996. Both the total percent change and the estimated annual percent change for incidence were based on incidence data from the five geographic areas for which data are available for each of three time periods, around 1950, 1969-71 and 1973-74 to 1996. Due to the limited availability of incidence data from the early time periods and the change in the composition of the non-white population over time, the incidence trends are presented for whites only. The estimates for children are for children of all races combined in Connecticut only. Mortality data are for the total United States and are for whites only for comparability to the incidence data. The last two columns display five-year relative survival figures for patients diagnosed during two time periods, 1950-54 and 1989-95 and are based on information from the End Results program for 1950-54 and SEER for 1989-95.

Caution should be exercised when interpreting these statistics. Evaluating trends over such a long period of time may hide recent changes in the trends. In addition, the straight line model fit to the log of the incidence and mortality rates and used to calculate the estimated annual percentage change may be inappropriate if the trend has changed directions or if the rate of change in rates has changed dramatically.

YEARS OF LIFE LOST DUE TO PREMATURE DEATH FROM VARIOUS CAUSES

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

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

CANCER PREVALENCE

There are different ways to define cancer prevalence. It could be the number of people who currently have cancer. It could also be the number of people who have ever had a particular cancer. Long-term incidence and survival rates from the State of Connecticut back to 1940 were used to estimate age-specific prevalence rates for Connecticut for a recent year. The age-specific prevalence rates for Connecticut were applied to the total U.S. population to estimate the number of Americans who were (will be) alive at a specific point in time who were ever diagnosed with invasive cancer (but including in situ bladder cancer). These prevalence estimates are an attempt to quantify the number of persons in the U.S. who have ever had a diagnosis of cancer (i.e., history of cancer). Prevalence estimates in this section were calculated based on Feldman (1986). There are several studies currently underway to evaluate different methods of calculating prevalence and the reliability of using data from 9 SEER areas back to 1973 or data only for the State of Connecticut back to 1940. Caution should be used in interpreting prevalence estimates.

PROBABILITY OF BEING DIAGNOSED WITH OR DYING FROM CANCER

Each site-specific section of the book contains a table with, the probability (expressed as a percent) of an individual of specified age being diagnosed with the specified cancer within ten, twenty or thirty years and within their total remaining lifetime. Lifetime risks of being diagnosed with cancer and lifetime risks of dying from cancer also appear (as percents) in each table and 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 1994-96 incidence and mortality rates from the SEER areas to a hypothetical cohort of individuals. This hypothetical cohort, consisting of an arbitrarily specified number of live births (e.g., 10,000,000), is considered at risk for two mutually exclusive events: 1) developing the specified cancer; and 2) death due to other causes without the specified cancer. Thus a standard multiple decrement life table is derived (with five-year age intervals up to age 94 and a 95+ interval) using these two types of events. In each age interval we start with the number alive and free of the specified cancer at the beginning of the interval, and subtract out the number who develop the specified cancer and the number who die of other causes among the cancer free. The lifetime risk of being diagnosed with the specified cancer is derived by summing all cancer cases from age 0 through 95+ and dividing by 10,000,000. This calculation does not assume an individual lives to any particular age, rather it is the sum over all age intervals of the probability of living to the beginning of each age interval times the probability of developing cancer in that interval. The probability of developing cancer during any time period (e.g., within 10 years of turning 50 years of age) 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.

For more details on this methodology see Feuer et. al (1992) and Feuer et al. (1993). One improvement over past calculations of the risk estimates was made in the population figures for people over age 85. To improve the precision of our calculations, populations for the age groups 85-89, 90-94, and 95+ were obtained by partitioning the 85+ figure from the SEER areas by interpolation using figures from the 1980 and 1990 decennial censuses. The BOC provided populations for these age groups for 1990 to 1997.

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, 1980) where a person is exposed to the risk of dying from the specified cancer and all other causes based on mortality data from the SEER registry areas. Although the lifetime risk of dying from cancer could have been derived for the entire U.S., these estimates were based only on data from SEER areas to allow comparison with the risk of diagnosis estimates.

U.S. CANCER MORTALITY RATES BY STATE

Each site-specific section of the book presents average annual mortality rates for the most recent 5-year period for all races by sex for selected cancers for all 50 states and the District of Columbia. The rates are per 100,000 and age-adjusted to the United States (U.S.) 1970 standard million population. The five states with the highest rates and the five states with the lowest rates are identified. The states are also ranked from highest rate to lowest rate for each of the cancers for which rates are reported. The percent difference (PD) between the individual state rates and the rate for the total U.S. is given and is based on the following formula:

$$PD = 100(\text{State rate} - \text{Total U.S. rate})/\text{Total U.S. rate}$$

The standard error provided for each age-adjusted rate is calculated, based on the assumption that, for each age-specific rate, the number of deaths is a Poisson random variable (Keyfitz, 1966) with the variance of the age-adjusted rate being a linear combination of the variances of the age-specific rates (Snedecor, 1980a). The difference between each age-adjusted state rate and the age-adjusted total U.S. rate is also tested for statistical significance by calculating a Z statistic from the following formula:

$$Z = (\text{State rate} - \text{Total U.S. rate})/SE_d$$

It is recognized that the two rates being compared are not independent because each state is part of the U.S.; however, this should not compromise the statistical test since each state represents a small proportion of the total U.S.

The standard error of the difference between two age-adjusted rates (SE_d) is given by the following formula:

$$SE_d = [(SE_s)^2 + (SE_u)^2]^{1/2}$$

where SE_s and SE_u are the standard errors of an individual state rate and the total U.S. rate respectively. The variance of each rate, i.e, the square of the standard error, is based on the Poisson assumption.

The standard error does not represent the total error which may be present in the age-adjusted rate but is merely the variance associated with the rates. In addition to this variance, there also exist potential biases and errors in the measurement of the rate which are extremely difficult to assess accurately and probably have a differential impact on the error for one state rate versus another.

Errors in the "measurement" of death rates can occur in either the numerator (the number of reported deaths) or the denominator (the population at risk). Sources of numerator error may include the under registration of deaths. Although investigation by the National Center for Health Statistics indicates that over 99% of all deaths in this country are registered, little is known concerning differentials by geographic area, age, sex, or race.

Numerator error also can occur due to misclassifications. These may include misclassification of race or ethnicity and cause of death. Recent research indicates that, for infant mortality, misclassification is highest for races other than white or black (Hahn, 1992). The extent of racial or ethnic misclassifications in death certificate coding, however, remains unknown.

In coding overall cancer mortality, misclassifications of cause of death would occur in those cases where the true cause of death was cancer, but a cause other than cancer was coded (and the reverse). Within the subset of all cancer deaths, there is the additional problem of misclassification of the primary cancer. It is already known, for example, that this is a problem with primary liver cancer (Percy, Ries, et al, 1990).

Denominator errors arise through census under- and over-enumeration in the decennial census (which is the base for intercensal population estimates and population projections). To the extent that any over- or under-count is substantial and variable among subgroups or geographic areas, it may have important consequences on death rates. The effect of an under-count is that it decreases the denominator leading to an over-estimation of the true rate. Conversely, an over-count would result in an under-estimation of the true rate.

In 1980, under-enumeration varied by age group with the greatest difference found for those 80 and older, who were under-counted by about five percent (U.S. Bureau of the Census, 1986). All other age groups were either over- or under-counted by less than 3 percent. For age-sex-race groups, the coverage was lowest for black men aged 40-49 where the under-count was 19 percent. It is thought that no improvement was achieved with the 1990 census, and in some instances, under-enumeration may be even worse than 1980.

The impact of any of these errors is that they alter the counts in either the numerator or the denominator, which in turn affect the calculated rate. Since the types of error encountered may differ by type of cancer, age group, race, sex, or even state, their impact is difficult to ascertain. Caution is recommended when dealing with those areas where potential problems may be present.

In testing the differences between the total U.S. rate and the rate for each state (and the District of Columbia) for a given cancer, it was necessary to consider the large number of statistical tests that were performed, because it would be expected that some tests are significant due to chance alone. To account for multiple comparisons, the overall significance level was chosen such that the probability that at least one comparison would be significant is 0.01. Furthermore, based on one of Bonferroni's inequalities (Snedecor, 1980b), the significance level for each individual comparison was set equal to 0.01/51, where 51 is the number of comparisons made for each type of cancer. Thus, any individual comparison with an associated p value less than 0.0002 was considered to be statistically significant.

Caution must be exercised in assessing statistically significant differences. Some states may have rates that are very close to the total U.S. rate, but because of their large population, the difference between their rate and the total U.S. rate is found to be statistically significant. On the other hand, some smaller states may have rates that differ substantially from the total U.S. rate, but because of their relatively small population, the differences are found to be statistically nonsignificant.

If the percent difference between the two rates is small, there may be some question as to the importance of the difference. It is difficult to specify a percent difference below which there would be no concern because the relative difference observed will be a function of the magnitude of the rates involved. It may also be of value to consider the size of the absolute difference between a state rate and the national rate in assessing the importance of a statistically significant difference. To further assist in the interpretation of the data, the tables are footnoted to indicate absolute differences greater than 15 percent, depending on the magnitude of the cancer rates.

It is important to note that comparing individual state rates with the total U.S. 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 is found to have the highest mortality rates. It can be argued that it is inappropriate to compare cancer rates for the District of Columbia with those from the 50 States because the District of Columbia is a predominantly urban area whereas states are comprised of a combination of urban, suburban, and rural areas. Mortality rates for many cancers are higher in urban areas. Also, the District of Columbia has a higher percentage of blacks (about two-thirds) than any state, and their higher mortality rates for several types of cancer elevate the overall rate for the District of Columbia.

INCIDENCE AND MORTALITY TIME TRENDS

Graphs depicting time trend lines are included for most of the individual cancers. Trend curves were fit using polynomial regression of the form

$$Y_x = B_0 + B_1x + ... + B_nx^n,$$

where Y_x is the rate in year x.

First order polynomials fit a linear trend representing either a constant yearly increase or decrease or a trend which is basically flat over the years involved. Second order polynomials may fit a trend whose function may increase or decrease to some maximum or minimum point in time before changing direction or whose rate of increase or decrease may not be constant. Polynomials higher than second order fit trend functions which may reach several maximum or minimum points. The correct function is determined by whether the addition of a higher order leads to a significant value for the coefficient, B, associated with that particular order (e.g., a trend is second order if and only if the term B_2 is significantly different from zero). Trend lines were not fitted for some cancers when annual rates showed substantial variation due to small numbers of cases in the numerator. Most of the trends were fitted with either first or second order polynomial functions.

INTERPRETATION OF CANCER STATISTICS

In reviewing the various cancer incidence, mortality, and survival statistics provided in this report, the reader should 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 is changing over time as the incidence of some cancers increases and the incidence of others decreases. Thus, the relative contribution of a specific cancer to the survival rate for all cancers combined may not be constant over time. Because survival rates differ by form of cancer, the overall cancer survival rate can fluctuate even when the survival rates for individual cancers remain unchanged. It is possible to adjust the survival rates for all cancers combined for a calendar period based on the relative frequency of each cancer for some specified reference period; however, rates adjusted in this manner have been found to 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: A factor that may lead to an artifactual increase in patient survival as well as incidence for a specific cancer is the detection and diagnosis of cancers earlier than otherwise expected. These changes can occur subsequent to the introduction of a new procedure to screen subgroups of the population for a specific cancer and need not be related to whether or not 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 will appear to increase. The additional survival associated with the time between a cancer being diagnosed by a screening procedure and the time at which the cancer would have been diagnosed in the absence of screening has been termed "lead-time" (Zelen, 1976) and results in an artifactual increase in patient survival. Screening for breast cancer has been demonstrated to result in increased survival over and above that resulting from "lead-time" alone. Screening for breast cancer has been demonstrated to reduce breast cancer mortality. The benefit of screening is being studied for some other cancers. Screening may also result in a decrease in survival rates for invasive cancer if the screening procedure consistently detects a cancer in a preinvasive phase. In this case, length-biased sampling (Zelen, 1976) may be operating and, if so, will result in those cancers that would have had a relatively good prognosis had they progressed to invasive disease being preferentially detected in a preinvasive phase. There is, therefore, the possibility of a systematic elimination of invasive cancers that would have had a relatively good prognosis. If this occurs, the mix of cancers that are not detected at screening and do progress to invasive becomes less prognostically favorable, resulting in a temporal decrease in survival for patients with invasive cancers. This latter effect of screening on patient survival may at least partially explain survival trends for cervical cancer. Other possible cancers affected include breast, colon, rectum and prostate.

Changes in diagnostic criteria: Early detection of cancer resulting from screening and/or earlier response to symptoms may result in the increasing diagnosis of small (early) tumors prior to their becoming 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 some of the cancers most likely to be affected.

Technological advances in diagnostic procedures: Temporal trends in survival for patients with specific cancers by stage at diagnosis, as well as temporal trends in distributions of stage at diagnosis, are not presented in this report. However, it is possible that the reader might compare survival 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 probability that a patient's cancer will be assigned to a particular stage may change over time due to advances in diagnostic technology. Utilization of new technology can give rise to a temporal phenomenon known as stage migration. Stage migration occurs when diagnostic procedures change over time, resulting in an increase in the probability that a patient's cancer will be diagnosed in a more advanced stage. For example, certain distant metastases which would have been undetectable a few years ago can now be diagnosed by a Computer Tomography (CAT) scan or by Magnetic Resonance Imaging (MRI). Therefore, some of the patients who would have been previously diagnosed as having cancer in a localized or regional stage would now be classified as having cancer in a distant stage. Thus, the likely result would be to remove the worst survivors, i.e,. those with previously

undetected distant metastases, from the localized and regional categories and put them into the distant stage category. As a result, the stage distribution for a cancer may become less favorable over time, but the survival rates for each stage category may improve. The latter occurs because those patients shifted from early to advanced stage likely have poorer survival than early stage patients, as indicated previously, but better survival than advanced stage patients as identified in past time periods. However, overall survival would not change. This has been referred to as the "Will Rogers phenomenon" (Feinstein, 1985) and is an important concept to understand when examining temporal changes in survival by stage as well as temporal changes in stage distributions. This phenomenon could affect staging for virtually all solid tumors.

Evolution of stage classifications: The American Joint Committee on Cancer has produced a new staging classification for many cancers every few years. The evolution of such classifications reflects the identification of new prognostic factors which may influence choice of treatment. Because the SEER Program collects data on extent of disease rather than some determination of stage specified in the medical record, changes in stage definitions can be more easily accommodated and trends in the new stage over time can be calculated if the detailed extent of disease has enough information to collapse to the new stage. For those cancers for which new prognostic variables are introduced into staging, such that previously collected detailed data on extent of disease cannot be collapsed into stage categories, there can be problems in assessing temporal trends in stage of disease. It is only possible to determine what effect changes in staging have had on stage-specific survival and stage distributions by reviewing the evolution of staging for a given cancer. One also needs to take stage migration (mentioned above) and extent of disease migration into account. One reason for using the historical categories of localized, regional and distant is that these categories have been fairly comparable over time.

Interpreting relative survival rates: The relative survival rate is the ratio of the observed survival rate to the expected survival rate for a patient cohort. When the population used in calculating the expected survival is similar to the cancer patients except for their cancer experience, the relative survival rate approximates the underlying cancer cause-specific survival. The expected rate is based on mortality rates for the total population taking into account, as appropriate, the age, sex, race, and calendar year of diagnosis of the patients. It is assumed that the presence of cancer is the only factor which distinguishes the cancer patient cohort from the general population, with the relative survival rate indicating the probability that patients will escape death due to causes associated with their diagnosed cancer. In some cases, there is a factor related to the risk of a cancer which is also 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, and therefore, a cohort of lung cancer patients will consist of a much higher proportion of smokers than the general population. However, smoking is a risk factor for other diseases resulting in smokers having a shorter life expectancy than non-smokers. Expected survival rates for lung cancer patients based on the general population will be unduly optimistic for this reason and will result in relative rates which are lower than they should be. The problem cannot be easily corrected because life-tables for smokers and non-smokers are not readily available. 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, 1991) which, in turn, may be related to life expectancy.

Previous to the CSR for 1973-1996, the expected rate tables used were for 1970 and 1980 and had separate tables for whites, blacks, American Indians, Chinese, Japanese, Filipinos, white Hispanic and Hawaiians. In updating the tables for 1990, several problems emerged. The U.S. lifetables are based on age, race and sex information from death certificates. The information on race on the death certificate may not be accurate since many times the funeral director will report race on the death certificate. Also, 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 mis-reporting exist across all races, they may be more problematic for races other than white or black because of their relatively smaller population sizes. Therefore, life tables were generated for 1970, 1980 and 1990 for only white, black and other and these lifetables were used to produce the relative survival rates in this book. Therefore, there may be small variations in survival rates calculated in this CSR to those in CSRs prior to 1973-1996.

Comparison with other databases: The SEER data are obtained from population-based cancer registries covering about 14 percent of the United States population. It is sometimes of interest to compare cancer

statistics for SEER areas with those from other registries both in the United States and worldwide. In making such comparisons, it is essential that the factors considered above be carefully considered for both data sources. In addition, completeness of case ascertainment, rules used to determine multiple primaries, follow-up, and rules used in assigning and coding cause of death should be assessed along with the sources and procedures used in obtaining population estimates. Depending on the rates being compared, there could be other confounding factors which should be adjusted for or otherwise considered. The same standard million should be used for the age-adjustment of each group being compared.

It is sometimes interesting to compare survival data for cancer patients in SEER areas with that from clinical trials. This must be done with great caution. Survival data from clinical trials may have been obtained from a patient population that is different from patients diagnosed in SEER areas in regard to prognostic factors for the cancer in question. 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. Patients from clinical trials may differ from patients diagnosed in SEER areas in regard to 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 the type discussed here.

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

Comparison of this report with previous reports: The cancer registries that participate in the SEER Program submit data on all cancers diagnosed in their coverage areas to the National Cancer Institute each year. Because of the dynamic nature of the registries' data bases, it is possible that the numbers of cancer cases in a particular race-sex-age-cancer category may change in a calendar year for which data have already been reported in a previous publication. One possible reason for this is that additional cancer cases that were previously overlooked for a given calendar year may be found and reported to the central registry. A second reason is 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 in a given year. A third reason is possible code changes that 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. A fourth reason is the elimination of duplicate records for the same patient, often due to name changes or misspellings.

This discussion has addressed issues that may result in a recent report having a different number of cases for a given time period than in an earlier report with its resulting effect 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 which are regularly updated by the Bureau of the Census. For example, previous population estimates for the nine years following the 1980 census were replaced with improved, new estimates controlled to population counts available from the 1990 census. Such changes may result in some differences between incidence and mortality rates for a calendar period as published in two different reports.

STANDARD ERRORS OF RATES

Survival rates: In the tables presenting survival rates, the reliability of the rates is indicated based on the magnitude of the standard error. In addition, if there were fewer than 25 total 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 noted in the footnote.

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

$$SE(CR_t) = CR_t \bullet \left[\frac{q_1}{e_1 - d_1} + \frac{q_2}{e_2 - d_2} + \ldots + \frac{q_t}{e_t - d_t} \right]^{1/2}$$

where CR_t is the t year relative survival rate, q_1 is the probability of dying in year 1, e_1 is the effective number of patients at risk in year 1, and d_1 is the number of deaths in year 1. The subscripts 2 through t refer to subsequent years 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 standard error of a particular incidence or mortality rate by the following formula for the standard error of a crude incidence or mortality rate (Keyfitz, 1966):

$$SE(rate) = rate/[events]^{1/2}$$

where events refer to the number of cancer diagnoses associated with an incidence rate or the number of deaths associated with a mortality rate.

Appendix Tables A-1 and A-2 provide numbers of cancer diagnoses within SEER and numbers of deaths in the total U.S., respectively, by race and sex for the most recent five-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 for a rate for a single year, the number of events is the number of diagnoses or deaths divided by five.

REFERENCES

American Cancer Society, Cancer Facts & Figures - 1999, Pub. # 99-300M, No. 5008.99, Atlanta, Georgia, 1999.

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.

Ederer F, Axtell LM, Cutler SJ. The relative survival rate: A statistical methodology. Natl Cancer Inst Monogr 1961; 6:101-121.

Elandt-Johnson RC, Johnson NL. Survival Models and Data Analysis. New York: Wiley, 1980.

Feldman AR, Kessler L, Myers M, Naughton MD. The prevalence of cancer: estimates based on the Connecticut Tumor Registry, New England Journal of Medicine, 315:1394-1397 (November 27), 1986.

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 Eng J Med 1985; 312:1604-1608.

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.

Feuer EJ, Wun L-M, Boring CC. Probability of developing cancer. In: Miller BA, Ries LAG, Hankey BF, Kosary CL, Edwards BK, eds. Cancer Statistics Review: 1973-1989, National Cancer Institute, NIH Pub. No. 92-2789, 1992, XXX.1-8.

Hahn RA, Mulinare J, Teutsch SM. Inconsistencies in coding of race and ethnicity between birth and death in U.S. infants. J Amer Med Assoc 1992; 267:259-263.

Keyfitz N. Sampling variance of standardized mortality rates. Hum Biol 38: 309-317, 1966.

Kleinbaum DG, Kupper LL, Muller KE. Applied Regression Analysis and Other Multivariable Methods. North Scituate, Massachusetts: Duxbury Press, 1988: 266-268.

Percy C, Van Holten V, Muir C. International classification of diseases for oncology, second edition, World Health Organization, Geneva, 1990.

Percy C, Ries LAG, Van Holten VD. The accuracy of liver cancer as the underlying cause of death on death certificates. Public Health Reports 1990; 105:361-368.

Snedecor GW, Cochran WG. Statistical Methods, Seventh Edition, The Iowa State University Press, 1980(a):188-189.

Snedecor GW, Cochran WG. Statistical Methods, Seventh Edition, The Iowa State University Press, 1980(b):115-117.

U.S. Bureau of the Census. Current Population Reports; Series P-25 No. 985. Washington: U.S. Government Printing Office., 1986.

Zelen M. Theory of early detection of breast cancer in the general population. In Breast Cancer: Trends in Research and Treatment. Heuson JC, Mattheiem WH, Rozencweig M, eds. New York, 1976: 287-299.