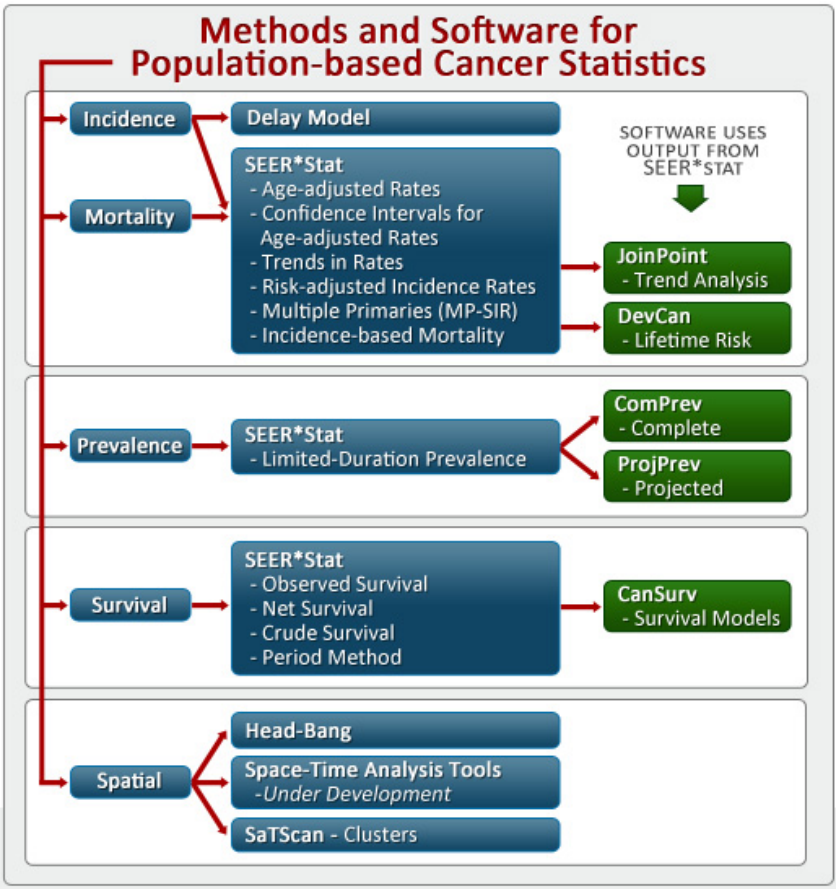
National Cancer Institute

Statistical Methods in the Cancer Surveillance Research Program (SRP)

Methods and Tools for Population-Based Cancer Statistics

SRP has developed new statistical methods and associated software tools for the analysis and reporting of cancer statistics. Different methods and software are available for calculating incidence, mortality, survival, prevalence, and spatial statistics. SEER*Stat contains a suite of tools for the analysis of the Surveillance, Epidemiology, and End Results (SEER) and other cancer-related databases. Methods associated with the reporting of basic cancer statistics are added directly to SEER*Stat. Methods involving complex modeling are developed as separate applications, and several can read data generated from SEER*Stat.

This document provides an overview of the methods and software available for the analysis of different types of cancer statistics.



SEER*Stat software has various session types, each designed for specific calculations:

1. Frequency session: Generates the number of records stratified by any variable in a database;
2. Rate session: Calculates disease incidence and mortality rates;
3. Survival session;
4. Prevalence session;
5. MP-SIR statistics session;
6. Case listing session: Allows users to view the values of variables for individual cases (records).

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

In all sessions, users must set options on tabs shown in the program's interface.

Methods for the Analysis of Incidence and Mortality Rates

- **SEER*Stat Rate session:** The following methods associated with the reporting of basic cancer incidence and mortality can be accessed from SEER*Stat. For more information and references, see *http://surveillance.cancer.gov/software/inc_mort.html*.

- **Age-Adjusted Rates:** An age-adjusted incidence or mortality rate is a weighted average of age-specific incidence or mortality rates in which the weights are the proportions of persons in the corresponding age groups of a standard million population. The potential confounding effect of age is reduced when comparing age-adjusted rates that are computed using the same standard population. Currently, several sets of standard population data are included in SEER*Stat. For more information about calculating age-adjusted rates, go to *http://seer.cancer.gov/seerstat/tutorials/aarates/definition.html*.

- **Gamma Confidence Intervals for Age-Adjusted Rates:** This method provides confidence intervals with good statistical properties. This remains true even when the cancer is rare, the population is small, or the population of interest is very different from the standard.

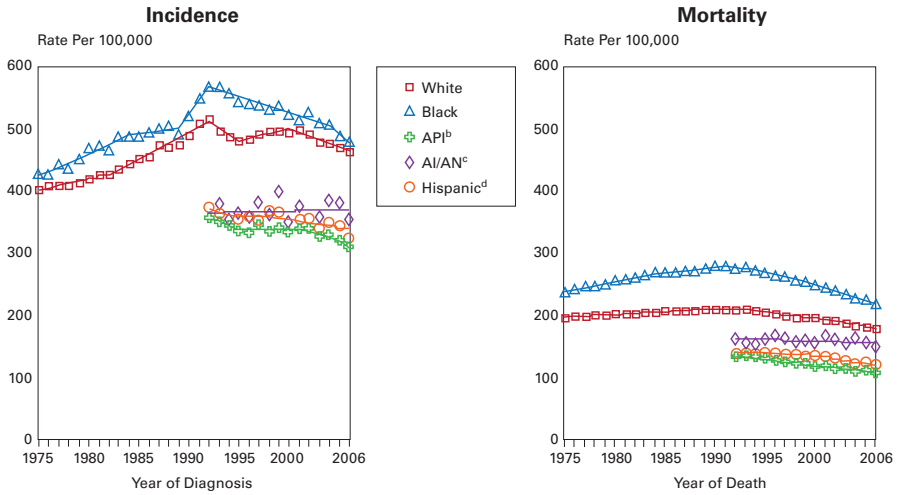
- **Trends in Rates:** Trends over time are based on frequencies (percent change, annual percent change, APC). The average APC is used to summarize trends or the change in rates over time.

- **Multiple Primary-Standardized Incidence Ratios (MP-SIR):** MP-SIR is a method used to perform multiple primary analyses and to test hypotheses that explore theoretical links in the etiology of two cancers. A defined cohort of persons previously diagnosed with cancer is followed through time to compare their subsequent cancer experience(s) to the number of cancers that would be expected based on incidence rates for the general population. MP-SIRs can be calculated in SEER*Stat. For more information on this topic, go to *http://seer.cancer.gov/seerstat/mp-sir.html*.

- **Incidence-Based Mortality (IBM):** The IBM rate allows a partitioning of mortality by variables associated with cancer onset. More specifically, IBM helps researchers determine whether or not variables such as year of diagnosis, age at diagnosis, stage of disease at diagnosis, histology, and treatment contributed to onset and subsequent deaths related to particular cancers. Accurately measuring incidence-based mortality requires access to high-quality population-based cancer registry data and high-quality follow-up of cancer patients for vital status, including cause of death. Incidence-based mortality rates can be calculated in SEER*Stat. To learn more about IBM and to access SEER*Stat software, go to: *http://surveillance.cancer.gov/statistics/ibm*.

- **Joinpoint Regression Program:** Joinpoint is an independent statistical software for the *analysis of trends* using joinpoint models, which are fit using joined segments (usually on a log scale). Cancer trends reported in NCI publications typically are calculated by using this program to analyze cancer incidence and mortality rates. Sample Joinpoint analyses are shown at right.

  Joinpoint software uses trend data (i.e., cancer rates) and determines the number and location of joinpoints that best fit the data. Joinpoint software can read output from SEER*Stat software or other types of datasets. If the model is fit on a log scale, then each segment can be characterized by a constant APC of the rates over the segment. This model enables users to deter-



SEER Incidence Rates and U.S. Death Rates for All Cancer Sites

mine if, how often, and when rates have changed. To download and access the latest version of Joinpoint, go to *http://surveillance.cancer.gov/joinpoint*.

- **DevCan:** DevCan is a statistical software used to calculate the lifetime risks of being diagnosed with or dying from cancer. DevCan takes cross-sectional counts of incident cases from the standard areas of the SEER Program and mortality counts from data collected by the National Center for Health Statistics and uses them to calculate incidence and mortality rates using population estimates from census data for these areas. These rates then are converted to the probabilities of developing or dying from cancer using a lifetable approach for a hypothetical population.

  The latest version of DevCan supports rate editing, sensitivity analysis, user-defined databases, confidence intervals, and the ability to obtain risk estimates from any age to any age. To download and access the latest version of DevCan, go to *http://surveillance.cancer.gov/devcan*.

Methods for the Analysis of Prevalence Rates

- **SEER*Stat Prevalence Session:** Prevalence represents new and pre-existing cases alive *on a certain date*, in contrast to *incidence, which reflects new cases of a condition diagnosed* during a given period of time. The prevalence session in SEER*Stat uses the counting method to estimate prevalence from incidence and follow-up data from the SEER Cancer Registries. Because the earliest SEER data used to estimate prevalence has incidence cases from 1975, the estimates are *limited duration prevalence* (*i.e., number or proportion of people alive on a certain day who had a diagnosis of the disease within the past "X" years*).

- **Complete Prevalence (ComPrev) Software:** ComPrev is a statistical software that calculates complete prevalence based on limited-duration prevalence statistics. Complete prevalence represents the proportion of people alive on a certain day who previously had a diagnosis of a disease, regardless of how long ago the disease (i.e., cancer) was diagnosed.

  ComPrev contains incidence and survival models estimated using SEER cancer data for a combination of cancer sites, sex, and races. It uses limited duration prevalence estimated from SEER*Stat to estimate the number of cases diagnosed prior to the registration period or prior to a given year. To learn more about ComPrev, go to *http://surveillance.cancer.gov/comprev*.

- **Projected Prevalence (ProjPrev) Software:** ProjPrev is a statistical software that takes limited-duration prevalence statistics from SEER*Stat and applies them to a different population. ProjPrev is used primarily to derive U.S. prevalence by projecting SEER prevalence onto populations in the United States. For more information about ProjPrev, go to *http://surveillance.cancer.gov/projprev*.

Methods for the Analysis of Survival Rates

- **SEER*Stat Survival Session:** Cancer survival statistics typically are expressed as the proportion of patients alive at some point subsequent to the diagnosis of their cancer. Various statistical methods and software tools have been developed for the analysis and reporting of cancer survival statistics.

Three measures of cancer survival can be calculated in the survival session of SEER*Stat software, including:

- ➢ **Observed all-cause survival:** Survival using all causes of death as an endpoint.
- ➢ **Net cancer-specific survival:** Cancer survival in the absence of other causes of death (the confounding effects of death from other causes are removed).
- ➢ **Crude probability of death:** Probability of death from cancer and the probability of death from other causes, each estimated in the presence of the other.

Cancer Survival Analysis Software (CanSurv): CanSurv is statistical software designed to model population-based survival data. For grouped survival data, CanSurv can fit both semi-parametric and parametric standard survival

models and mixture cure models. It also can fit parametric (cure) survival models to individually listed data. CanSurv uses population-based survival data extracted from SEER*Stat Survival sessions. For more information about CanSurv, go to *http://surveillance.cancer.gov/cansurv*.

Geographic Information Systems, Spatial Analysis, and Data Visualization

SRP has an active research program in the areas of geographic information systems (GIS), which includes geo-spatial database development, geo-statistical analysis of cancer-related data, and data visualization. In addition, SRP consults across the Division of Cancer Control and Population Sciences on the analysis and presentation of geographic data, and helps coordinate extramural geographic-based research in cancer control and epidemiology. More information about GIS technology and its various applications in NCI research can be found at *http://gis.cancer.gov*.

- **Geo-spatial database development:** The goal of this activity is to build an integrated database system to support research that examines patterns of cancer development and exposure assessment at multiple geo-spatial resolutions. Examples include:

  - Neighborhood-level database (census tract, county, state): This allows researchers to look at socio-demographic profiles, built environment, food and diet consumption, physical activity, and urban sprawl.

  - Tobacco control database: Designed for users who are interested in looking at tobacco policy coverage, smoking prevalence, tobacco taxes, and tobacco expenditures (county, health service area, state).

  - Area-level environmental exposure data: Ideal for studying UV, pollution, and toxins (fine scale).

- **Geo-spatial and statistical analysis:** The goal of this effort is to develop a modeling framework to better articulate relationships among environmental exposures, social risk factors, behaviors, new treatments, incidence, prevalence, mortality, and survival rates. Examples include:

  - Spatial-temporal prediction of the numbers of new cancer cases for every U.S. county and state, based on hierarchical statistical models of cancer patterns in SEER counties.

  - Methods to optimize cluster detection using scan statistics.

  - Small-area analysis to identify the various causes attributed to increasing incidence rates for certain types of cancers.

  - Spatio-temporal modeling of tobacco policy and its impacts on smoking prevalence.

  - Urban sprawl, obesity, and cancer mortality in the United States.

Tools for Geo-spatial and Statistical Analysis

- **Head-Bang PC Software:** "Head banging" is a weighted, two-dimensional, median-based smoothing algorithm developed to reveal underlying geographic patterns in data when the values to be smoothed do not have equal variances. Geographic smoothing algorithms "borrow information" from neighboring areas to stabilize results from sparsely populated areas. This reduces variability in the data and allows patterns to emerge, and also increases bias in estimates for smaller areas. Variance reduction also allows users to identify and compare clusters of counties with similar values. To learn more about Head-Bang PC software, go *http://surveillance.cancer.gov/headbang*.

- **Space-Time Analysis Tools** (under development): Space-time analysis tools are used to analyze peoples' motions through space and time. For example, these tools can track an individual's lifetime exposure while taking into account residential histories, or track an individual's access to cancer screening or treatment services while noting daily travel. Currently, SRP is working on two space-time analysis tools:

  - A tool that facilitates analysis of space-time paths (i.e., both residential histories and daily travel).

  - An exploratory/spatial/visualization/temporal tool that displays multivariate data by both space and time.

  To learn more about SRP's upcoming space-time analysis tools, go to *http://gis.cancer.gov/nci/geovisualization.html*.

- **SatScan (Spatial and Space-Time Scan Statistics):** SatScan software analyzes spatial, temporal, and space-timepoint data using the spatial, temporal, or space-time scan statistic. SatScan is designed for any of the following purposes:

  - To evaluate reported spatial or space-time disease clusters to determine if they are statistically significant.

  - To test whether a disease is randomly distributed over space, over time, or over space and time.

  - To perform geographical surveillance of disease to detect areas of significantly high or low rates.

  - To perform repeated time-periodic disease surveillance for the early detection of outbreaks.

  To download and access the latest version of SatScan, go to *http://surveillance.cancer.gov/satscan*.

- **Visualization Tools:** As cancer rates and patterns are better understood through data collection and analysis, it becomes more important to effectively represent this information in a way that is useful to and interpretable by various audiences. The goal of the projects on visualization tools is to develop innovative research and technologies to display geo-referenced data and statistics. Examples in this area include:

  - **State Cancer Profiles:** A Web-based tool for mapping, query, and dissemination of cancer and risk factor statistics, primarily for public health professionals (*http://statecancerprofiles.cancer.gov*)

  - Linked Micromaps: A multivariate visualization tool *(http://gis.cancer.gov/tools/micromaps)*

Survey Methods, Design, and Analysis

Sample surveys are a vital tool in cancer surveillance and control. SRP has provided statistical collaboration and support for the design and analysis of sample surveys and has participated in the development of new methods.

Examples of completed and ongoing projects include: (1) statistical design support (i.e., sample size determination, power analysis, weighting, and imputation) for numerous surveys such as the Health Information National Trends Survey (HINTS), the Tobacco Use Supplement to the Current Population Survey (CPS-TUS), the California Health Interview Survey (CHIS), and County-level Estimates of Smoking, etc.; (2) providing statistical analysis assistance (statistical modeling and analysis of trends) to researchers for numerous studies utilizing data obtained from a complex survey design; and (3) development of novel statistical techniques for improved small-area estimates of cancer risk factors or control activities (e.g., screening rates) using data from a single survey such as the CPS-TUS, or combined data from multiple surveys such as the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System (BRFSS). For further information, visit *http://sae.cancer.gov*.

Simulation Modeling To Guide Public Health Research and Priorities

SRP sponsors a cooperative group of grantees who utilize modeling to investigate the impact of interventions (screening, treatment, primary prevention) on population-based cancer trends in the United States for breast, prostate, colorectal, esophageal, and lung cancer. This group is collectively known as the Cancer Intervention and Surveillance Modeling Network (CISNET). More information about CISNET is available at *http://cisnet.cancer.gov*.

Methodology To Estimate Cancer Incidence Rates Adjusted for Reporting Delay

Reporting Delay: Reporting delay is defined as the time that elapses before a diagnosed cancer case is reported to NCI. Because the collection, collation, and correction of data is a complex process, there is some delay from the time that a case is diagnosed until it is reported centrally. Current NCI guidelines allow for a standard delay of 22 months (or close to 2 years) between the end of a diagnosis year and the time cancers are first reported to NCI. The data then are released to the public in the spring of the following year. In future submissions, prior years of data will be updated if new cases are found, or corrections are made for existing cases.

Reporting delay is used to adjust a current case count to account for anticipated future corrections (both additions and deletions) to the data. Adjusted counts and associated delay models are beneficial in that they allow for precise determination of current cancer trends, monitoring the timeliness of ongoing data collection, and quality control. For more information on the delay adjustment method, go to *http://surveillance.cancer.gov/delay*.

Publications

For a complete listing of SRP publications, visit *http://surveillance.cancer.gov/publications*.

Funding Opportunities

For funding opportunities in biostatistics, visit *http://statfund.cancer.gov*. The site also is a resource on statistical grants and application procedures.

Employment and Training Opportunities

The Statistical Methodology and Applications Branch welcomes applications from interested individuals. For a listing of currently open positions, visit *http://surveillance.cancer.gov/jobs*.