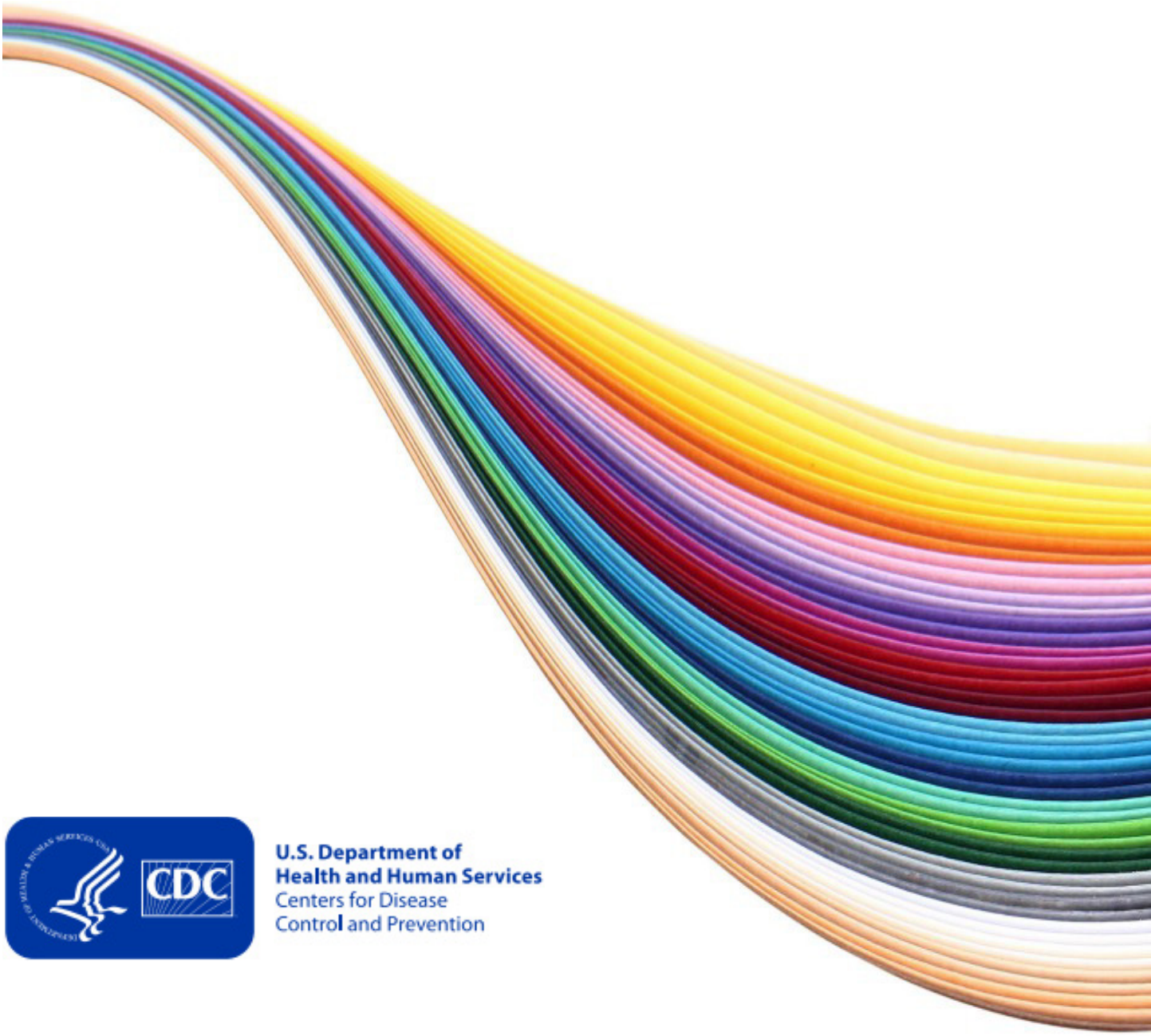
U.S. Cancer Statistics Public Use Database Technical Documentation

U.S. Data
November 2021 Submission
Diagnosis Years 2001–2019



U.S. Department of Health and Human Services
Centers for Disease Control and Prevention

Table of Contents

U.S. Cancer Statistics Public Use Databases

Researchers can access and analyze high-quality population-based cancer incidence data on the *entire* United States population.

De-identified cancer incidence data reported to CDC's National Program of Cancer Registries (NPCR) and the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) Program are available to researchers for free in public use databases that can be analyzed using software developed by NCI's SEER Program.

Cancer surveillance data from CDC and NCI are combined to become U.S. Cancer Statistics, the official source for federal cancer data. U.S. Cancer Statistics public use databases include cancer incidence and population data for all 50 states, the District of Columbia, and Puerto Rico, providing information on more than 33 million cancer cases.

Documentation for U.S. Data (2001–2019)

Two United States Cancer Statistics public use databases are available for researchers: the U.S. (2001–2019) database, described in this section, and the U.S. and Puerto Rico (2005–2019) database.

The U.S. (2001–2019) database—

- Includes race and ethnicity variables.
- Does not include Puerto Rico data.
- The population denominators are race-specific, ethnicity-specific, and sex-specific county population estimates from the U.S. Census (July 1, 2010–2020 bridged–race vintage 2020 population estimates), modified by SEER and aggregated to the state and national levels.

Population Coverage by Diagnosis Year

For cases diagnosed from 2003 through 2017, 100% of the population is covered for all 50 states and the District of Columbia. In 2001 and 2002, cases that were diagnosed in Mississippi are not available. In 2018 and 2019, cases that were diagnosed in Nevada are not available. The U.S. population covered for each of those four years is 99%. The U.S. population coverage for 2001 through 2019 is 98%.

Suggested Citations

Please use these standard citations for tables and figures when presented in presentations or publications.

For population coverage: Data are from population-based registries that participate in CDC's National Program of Cancer Registries and/or NCI's Surveillance, Epidemiology, and End Results Program and meet high-quality data criteria. These registries cover approximately [XX]% of the U.S. population.

For age-adjusted rates: Rates are per 100,000 persons and are age-adjusted to the 2000 U.S. standard population (19 age groups – Census P25–1130).

For the database: National Program of Cancer Registries and Surveillance, Epidemiology, and End Results Program SEER*Stat Database: NPCR and SEER Incidence – U.S. Cancer Statistics 2001–2019 Public Use Research Database, 2021 submission (2001–2019), United States Department of Health and Human Services, Centers for Disease Control and Prevention and National Cancer Institute. Released June 2022. Available at www.cdc.gov/cancer/uscs/public-use.

Archived Documentation

- U.S. Cancer Statistics 2001–2018 Public Use Database Data Standards and Data Dictionary [PDF-490KB]
- U.S. Cancer Statistics 2001–2017 Public Use Database Data Standards and Data Dictionary [PDF-467KB]
- U.S. Cancer Statistics 2001–2016 Public Use Database Data Standards and Data Dictionary [PDF-1MB]

- U.S. Cancer Statistics 2001–2015 Public Use Database Data Standards and Data Dictionary [PDF-890KB]
- U.S. Cancer Statistics 2001–2014 Public Use Database Data Standards and Data Dictionary [PDF-978KB]
- National Program of Cancer Registries (NPCR) Public Use Research Data — Data Standards and Data Dictionary (diagnosis years 2001–2013) [PDF-1.3MB]

Cautionary Notes for U.S. Data

Note: Before using the U.S. (2001–2019) data, read and understand the following information. If you have questions, please contact CDC at uscsdata@cdc.gov.

Case Inclusions and Exclusions

Cancer registries that are supported by CDC's National Program of Cancer Registries (NPCR) or the National Cancer Institute's (NCI's) Surveillance, Epidemiology, and End Results (SEER) program report all incident cases coded as *in situ* (non-malignant), invasive (malignant; primary site only), and non-malignant (including borderline and benign) central nervous system tumors according to the *International Classification of Diseases for Oncology, Third Edition* (ICD-O-3), with the following exceptions—

- *In situ* cancers of the cervix are not reported.
- Basal and squamous cell carcinomas of the skin are not reported, except when these occur on the skin of the genital organs.
- *In situ* cancers of the urinary bladder are re-coded as invasive behavior because the information that distinguishes between *in situ* and invasive bladder cancers is not always available or reliable. Stage for these cases remains coded as *in situ*.[1]

Additionally, in these public use databases—

- Cases with an unknown age or with sex other than male or female have been excluded from the database. The frequency counts will not change based on whether *Known Age* or *Male or Female Sex* is checked on the SEER*Stat Selection tab.
- *Malignant Behavior* is a default selection for this database, as this restriction is used by CDC's NPCR and NCI's SEER Program for generating most official cancer statistics. Malignant behavior is defined by the variable *Behavior Code ICD-O-3*. This database includes *in situ* and nonmalignant central nervous system (CNS) cases. These nonmalignant cases can be analyzed by unselecting the *Malignant Behavior* check box on the SEER*Stat Selection tab.

Changes Made to the Database

In the June 2022 release, the following changes were made to the data—

- The variable describing race was revised and renamed to *Race recode (W, B, AIAN, API);* it was *Race recode for uscs* in previous years. The "other" and "unknown" categories were collapsed to one "Unknown" category.
- The *Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)* variable was added to describe both race and ethnicity.
- Delaware, Kentucky, and Pennsylvania no longer require race and ethnicity suppressions when data are presented by state. The *state race eth suppress* variable has been updated to reflect this change.
- Three changes were made to *Merged Summary Stage in this release*.

- Stage data for all testis cases diagnosed in 2018 and 2019 were excluded.
- Stage data for all myeloma and leukemia cases were excluded.
- If the behavior was coded as benign or borderline for brain and central nervous system cases, then *Merged Summary stage* were coded as benign/borderline.

Previous changes made to this database are described on the U.S. Data Change History page.

Suppression Rules[2 3]

Suppressing Fewer Than 16 Cases

The suppression rule is fewer than 16 cases for the time period based on rate stability. This suppression rule is applied automatically in these databases.

When the number of cases used to compute the incidence rates is small, those rates tend to have poor reliability. Therefore, to discourage misinterpretation and misuse of counts, rates, and trends that are unstable because of the small number of cases, these statistics are not shown in tables and figures if the counts are fewer than 16 for the time period. A count of fewer than about 16 in a numerator results in a standard error of the rate that is about 25% or more as large as the rate itself. Equivalently, a count of fewer than about 16 results in the width of the 95% confidence interval around the rate being at least as large as the rate itself. These relationships were derived under the assumption of a Poisson process and with the standard population age distribution close to the observed population age distribution.

Another important reason for employing a cell suppression threshold value is to protect the confidentiality of patients whose data are included in a report by reducing or eliminating the risk of identity disclosure. The cell suppression threshold value of 16 is recommended to protect patient confidentiality given the low level of geographic and clinical detail provided.

Complementary Cell Suppression

Complementary cell suppression prevents users from subtracting to find suppressed counts. Use this practice when any suppression occurs in the data presentation. In addition, when information from other cells, tables, or figures can be used to determine a suppressed cell, suppress at least one other cell. When analyzing data at the state or regional levels, suppress counts for national and regional data if a single state in a region is suppressed. Rates, confidence intervals (CIs), and populations can be shown at the national and regional levels. Rates, confidence intervals (CIs), and populations can be shown at the national and regional levels. Use this suppression when a single or multiple years of data are being presented.

Race and Ethnicity Suppression

States have the option to suppress race-specific and Hispanic ethnicity-specific data every submission year. While these states can be included in an aggregated analysis, the affected state's race and ethnicity information cannot be reported at the state level.

The merged system-supplied variable, *state race ethnicity suppress,* can be used to restrict your analysis to the states that are eligible to be included in a state-level analysis of race and ethnicity combinations. If conducting a state-level analysis of race or ethnicity only, manually make restrictions in the SEER*Stat Selection tab.

The following states have race or ethnicity data presentation restrictions—

- Data for American Indian and Alaska Native people cannot be displayed for Illinois, Kansas, New Jersey, and New York.
- Data for Asian and Pacific Islander people cannot be displayed for Illinois and Kansas.
- Hispanic ethnicity data cannot be displayed for Illinois and Massachusetts.
- Race and ethnicity combinations—White Hispanic, White non-Hispanic, Black Hispanic, and Black non-Hispanic—cannot be displayed for Illinois, Kansas, and Massachusetts.

For more information, please refer to the *Race recode (W, B, AIAN, API)*, *Origin recode NHIA (Hispanic, Non-Hisp)*, and *Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)* variable descriptions.

Case-Level Data

As a further mechanism to protect data confidentiality and due to data sharing agreements with some states, the case listing function in SEER*Stat has been disabled for this database.

Benign Central Nervous System (CNS) Tumors

Cancer registries began collecting information on nonmalignant brain and other central nervous system tumors with cases diagnosed in 2004. Collection of these tumors is in accordance with Public Law 107-260, the Benign Brain Tumor Cancer Registries Amendment Act, which mandates that NPCR registries collect data on all brain and other central nervous system tumors with a behavior code of 0 (benign) or 1 (borderline), in addition to *in situ* and malignant tumors. Data for nonmalignant brain and other nervous system tumors were available from all registries contributing to this report.

Behavior

The behavior variable in the current database is *Behavior Code ICD-O-3.* Previous database releases included the variable *Behavior Recode for Analysis.*

The database's default is to restrict analyses to malignant cases. This restriction is used by CDC's NPCR and NCI's SEER Program for generating most official cancer statistics. To analyze benign, borderline, or *in situ* cases, uncheck the "Malignant Behavior" box in the SEER*Stat Selection tab.

To create comparable analyses using a database with data from submission years 2018 and earlier–

- Uncheck the "Malignant Behavior" box in the SEER*Stat Selection tab.
- Add the following selection criteria: {Site and Morphology.Behavior recode for analysis} = 'Malignant','Only malignant in ICD-O-3','Only malignant 2010+'.

Primary Site Variables[4–8]

Beginning in diagnosis year 2010, some lymphoma and leukemia ICD-O-3 codes were updated based on changes from the World Health Organization. The appropriate site recode variables to include these updates are *Site recode ICD-O-3/WHO 2008* for all ages and *International Classification of Childhood Cancer (ICCC) site recode ICD-O-3/WHO 2008* and *ICCC site rec extended ICD-O-3/WHO 2008* for the childhood cancer recodes.

Consider reviewing the variable *Site recode ICD-O-3/WHO 2008* before using the directly coded primary site. See more information on the SEER primary site recodes.

Stage

A merged variable, *Merged Summary Stage,* is provided to span time periods when three different staging schemes are used. The coding logic for this merged variable is—

- For NPCR-registries—
  - If a case was diagnosed in 2001, 2002, 2003, 2016 or 2017, stage at diagnosis is recorded using the *SEER Summary Stage 2000* variable value.
  - If a case was diagnosed in or between 2004 and 2015, stage at diagnosis is recorded using the *Derived SEER Summary Stage 2000* variable value. If the *Derived SEER Summary Stage 2000* variable is blank or unstaged, and the *SEER Summary Stage 2000* variable has a valid value, that value is used to populate the merged variable.
  - If a case was diagnosed in 2018 or 2019, stage at diagnosis is recorded using the *Summary Stage 2018* variable value.

- For SEER-only registries (Connecticut, Hawaii, Iowa, and New Mexico)—
  - If a case was diagnosed in 2001, 2002, or 2003, stage at diagnosis is recorded using the *SEER Summary Stage 2000* variable value.
  - If a case was diagnosed in or between 2004 and 2017, stage at diagnosis is recorded using the *Derived SEER Summary Stage 2000* variable value.
  - If a case was diagnosed in 2018 or 2019, stage at diagnosis is recorded using the *Derived Summary Stage 2018* variable value.

Notes for users of this variable include—

- Due to changes made in the Summary Stage 2018 Coding Manual, for cases diagnosed in 2018 or 2019—
  - The category *Regional, NOS* (code 5) is no longer used.
  - There is an artificial increase in the category *Regional by Direct Extension Only* (code 2) for brain, CNS Other, and lymphoma cases. This is because *Regional, NOS* for these cases changed from code 5 to code 2.

- *Merged Summary Stage* data are not available for testis cases.

Reporting Delay[9]

NPCR and SEER registries annually submit all eligible years of data to CDC and NCI, respectively. As a result, cases submitted in previous years may be deleted, and new cases diagnosed in previous years may be added. The addition of new cases is called a *reporting delay*. This reporting delay may cause an appearance of decreasing trends. For example, reporting of melanoma cases diagnosed in an outpatient facility may be delayed. As a result, the trend in incident melanoma cases might superficially appear to have dropped in the most recent year.

Checking SEER*Stat Frequencies

You can check the setup of your SEER*Stat program by comparing results to those published in the U.S. Cancer Statistics Data Visualizations tool. Note that the data in the Data Visualizations tool are restricted to malignant behaviors. Be sure the Malignant Behavior box is selected in the SEER*Stat Selection tab.

References

[1]Young JL Jr, Roffers SD, Ries LAG, Fritz AG, Hurlbut AA (eds). *SEER Summary Staging Manual – 2000: Codes and Coding Instructions.* National Cancer Institute, NIH Pub. No. 01-4969, Bethesda, MD, 2001.

[2]Federal Committee on Statistical Methodology. *Report on Statistical Disclosure Limitations Methodology (Statistical Working Paper 22).* [PDF-745KB] Washington, DC: Office of Management and Budget; 2005.

[3]Doyle P, Lane JI, Theeuwes JM, Zayatz LM. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: Elsevier Science; 2001.

[4]Fritz A, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin D, et al., editors. *International Classification of Diseases for Oncology, Third Edition.* Geneva: World Health Organization; 2000.

[5]*International Classification of Diseases for Oncology, Third Edition, First Revision.* Geneva: World Health Organization, 2013.

[6]Ruhl J, Adamo M, Dickie L. (January 2015). *Hematopoietic and Lymphoid Neoplasm Coding Manual.* [PDF-806KB] National Cancer Institute, Bethesda, MD.

[7]Surveillance, Epidemiology, and End Results Program. *2007 Multiple Primary and Histology Coding Rules.* Bethesda, MD: US Department of Health and Human Services, National Cancer Institute; Revised August 24, 2012; Accessed January 25, 2017.

[8]Surveillance, Epidemiology, and End Results Program. Hematopoietic and Lymphoid Neoplasm Database. Bethesda, MD: US Department of Health and Human Services, National Cancer Institute; 2016.

[9]Clegg LX, Feuer EJ, Midthune DN, Fay MP, Hankey BF. Impact of reporting delay and reporting error on cancer incidence rates and trends. *Journal of the National Cancer Institute* 2002;94(20):1537–1545.

U.S. Data (2001–2019) Analyses Checklist

Multi-Year Analyses

The database includes variables that can be used to restrict analyses to the states meeting U.S. Cancer Statistics publication criteria during the most commonly analyzed multi-year time periods, specifically—

- All years of data in the database (variable *USCS0119* for diagnosis years 2001–2019).
- The most recent 10 years of data (*USCS1019* for diagnosis years 2010–2019).
- The most recent 5 years of data (*USCS1519* for diagnosis years 2015–2019).

If you are conducting a multi-year analysis and want to restrict it to the states that met publication criteria during each of the years, did you use variable *USCS0119, USCS1019,* or *USCS1519* and also use the *Year of Diagnosis* variable on the SEER*Stat Selection tab?

- This is important for trend analyses so same states need are included for each year.
- The *Year of Diagnosis* variable is used in combination with the predefined USCS variable to exclude the non-relevant years. For example, if *USCS1519* is used, then *Year of Diagnosis* should also be restricted to diagnosis years 2015–2019 in the SEER*Stat Selection tab.
- If you would like to analyze a range of years other than those predefined variables, please contact CDC at uscsdata@cdc.gov and we will create a new variable for you.

Single-Year Analyses

If you are analyzing just one year of data, did you use the variable *USCS Standard* and restrict the analysis to the specific *Year of Diagnosis* in the SEER*Stat Selection tab?

Common Selection and Reporting Considerations

- If you are reporting **state-level race, ethnicity or race/ethnicity combinations,** have you suppressed data from the registries that opted out of reporting these data items? Race and ethnicity combinations can be excluded using the *State Race Ethnicity Suppress* variable; race-only or ethnicity-only suppressions should be done manually in the SEER*Stat Selection tab.
- If a user-defined **primary site variable** was created (rather than using the *Site recode ICD-O-3/WHO 2008* variable)—
  - Did you exclude leukemias and lymphomas (9590–9992)?
  - Did you consider excluding Kaposi sarcoma (9140) and mesothelioma (9050–9055)?

  For more information, see the Primary Site Variables description.

1. If your analysis includes **histology,** and if appropriate for the cancer site, did you use the *Diagnostic Confirmation* variable to specify the analysis be limited to microscopically confirmed cases?

2. If you are analyzing **sex-specific cancers** (such as prostate cancer or female breast cancer), did you limit the analysis to the appropriate sex to get the correct population denominator?

3. When reporting **rates,** have you included the label "per 100,000 persons," "per 100,000 women," or "per 100,000 men"?

4. Have you included **citations** for the—
   - Percentage of United States population coverage provided by the database?
   - NPCR and SEER Incidence – U.S. Cancer Statistics 2001–2019 Public Use Research Database?

U.S. Data Variables

The following variables are available in the U.S. Cancer Statistics Public Use Database, U.S. data (2001–2019). They are listed by SEER*Stat category. Click on the variable name for more information, including the source, description, and considerations for use.

Age at Diagnosis

- Age recode with <1 year olds

Race, Sex, Year of Diagnosis, and Registry

- Sex
- Year of diagnosis
- Addr at DX – state
- USCS standard
- Race recode (W, B, AIAN, API)
- Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)
- Program
- Region
- USCS0119
- USCS1519
- Origin recode NHIA (Hispanic, Non-Hisp)

Site and Morphology

- Primary site – labeled
- Histologic type ICD-O-3 (International Classification of Diseases for Oncology, Third Edition)
- Behavior code ICD-O-3
- Grade
  Restricted to diagnosis years 2001–2017
- Grade clinical
  Restricted to diagnosis year 2018 or later
- Grade pathological
  Restricted to diagnosis year 2018 or later
- Diagnostic confirmation
- ICD-O-3 histology/behavior, labeled
- Site recode ICD-O-3/WHO 2008
- International Classification of Childhood Cancer (ICCC) site recode ICD-O-3/WHO 2008
- ICCC site recode extended ICD-O-3/WHO 2008
- Adolescent and young adult (AYA) site recode 2020
- Lymphoid neoplasm recode 2021

Stage – Local, Regional, Distant (LRD) [Summary and Historic]

- Merged summary stage
  Stage data are not available for testis cases.

Therapy

- Rx summary – surgery primary site
  Restricted to female breast and diagnosis years 2003 or later

Extent of Disease – Collaborative Stage (CS)

- CS site-specific factor 1 (WHO Grade Classification)
  Restricted to brain and central nervous system and diagnosis years 2011- 2017
- Laterality
- Merged estrogen receptor
  Restricted to female breast and diagnosis years 2004 or later
- Merged progesterone receptor
  Restricted to female breast and diagnosis years 2004 or later
- Merged HER2 summary
  Restricted to female breast and diagnosis years 2010 or later

Multiple Primary Fields

- Sequence number – central

Dates

- Year of birth
- Month of diagnosis

User-Specified

- Rural-urban Continuum 2013, grouped

Merged System-Supplied

- Alcohol-related cancers
- Human papillomavirus (HPV)-related cancers
- Obesity-related cancers
- Physical inactivity-related cancers
- Tobacco-related cancers
- State race ethnicity suppress