Description of CiNA Public Use Dataset

ACCESS to CiNA Public Use is through DaRT (Data Request Tracking) system accessible here: https://www.naaccr.org/cina-data-products-overview/

The NAACCR CiNA Public Dataset is updated with most recent diagnosis year annually in early summer. The goal behind producing the CiNA Public Use Dataset is to provide faster, streamlined access to the CiNA data in order to increase CiNA data use and, ultimately, reduce the burden of cancer. The data will be available annually no later than July.

The CiNA Public Use Dataset is a publically accessible, non-confidential data set with a limited number of variables, available in the SEER*Stat program. Access requires only a signed Data Assurances Agreement for access. There are other CiNA databases with more extensive variable set that require a NAACCR approval and a “yes” consent by each participating registry.

The CiNA Public Use variable list is included at the end of this document. Many variables are recoded from the reported data for ease of use and standardization of analysis. There are no treatment variables available in the CiNA Public Use Dataset.

The dataset is released through the SEER*Stat system and a user may generate counts, rates and tends. This dataset includes age in the 19 age group categories. No case-level data is exportable from CiNA Public Use. SEER*Stat provides automatic cell suppression of <6 and scrambles the Patient IDs.

There are two CiNA Public Use Datasets. The NAACCR CiNA Public Use standard file uses 20-year age groups for age-adjusting. US registries have diagnosis years 2000-2020 available, with the exception of Puerto Rico. Canada has diagnosis years 2001-2020, and Puerto Rico has diagnosis years 2010-2020 available. A user can also request the old, 19 age group Public Use file, starting with diagnosis year 1995 for all eligible registries.

Additional details about the CiNA Public Use Dataset are provided in this document. If you have questions about the NAACCR CiNA Public Dataset, please contact Recinda Sherman, Manager of Research and Data Use at rsherman@naaccr.org or 217-698-0800 x6.

Citation

Please reference to the source of these data in any published document as indicated in SEER*Stat session.
For example:
NAACCR Incidence - CiNA Public File, 1995-2018 (which includes data from CDC’s National Program of Cancer Registries (NPCR), CCCR’s Provincial and Territorial

Registries, and the NCI's Surveillance, Epidemiology and End Results (SEER) Registries), North American Association of Central Cancer Registries.

Technical Documentation

The NAACCR CiNA-Public Dataset is distributed through SEER*Stat and contains individual records of cancer incidence among US and Canada residents diagnosed from 1995 to the most current year of diagnosis.

The purpose of releasing cancer surveillance data is to inform public health decision making. Cancer rates are often needed for subgroups or for small populations in order to understand the burden of cancer in these groups or areas. But working with small numbers has two problems 1) working with small numbers, particularly linking with external data, has the potential for confidentiality breaches; and 2) small numbers raise statistical issues regarding the accuracy and, ultimately, the usefulness of the data.

- To preserve confidentiality of the data, data will be automatically suppressed for counts less than 6 based on potentially linkable variables (registry, sex, age, race, race/ethnicity, year of diagnosis and site).
- For issues of statistically stability, we advise caution in interpreting rates and other results based on fewer than 25 cases.

Software

SEER*Stat statistical software is a standard tool for analysis of cancer-related data. SEER*Stat is distributed with the CiNA Public Dataset. Additional information on SEER*Stat is available on the NCI, SEER site: http://seer.cancer.gov/seerstat/. Tutorials are available here: http://seer.cancer.gov/seerstat/tutorials/. Delay factors, survival statistics, and prevalence are not currently available for the CiNA Public Dataset.

Representation

To be included in the CiNA Public dataset, a central registry from the US or Canada must meet specific data quality standards. All Gold and Silver NAACCR-certified central registries are eligible for inclusion in the CiNA Public dataset. Each central registry must also consent to the use of their data in the CiNA Public dataset. A current list of certified registries is available here: https://www.naaccr.org/certified-registries/. Registries may have not been certified in prior years, but if their data quality improves over time, their data is included in CiNA. However, not all states meet the data quality criteria for each year and will have zero counts for those years. Please review *Registry Data Fitness for use By Data Year* available here: https://www.naaccr.org/cina-data-products-overview/ .

Data Collection

Cancer registry data is collected in an on-going, systematic, and standardized process. In Canada, the cancer registry collection program is overseen by the Canadian Council of Cancer Registries. In the US, there are two cancer registry collection programs—the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER Program) and the Center for Disease Control's National Program of Cancer Registries (NPCR). Data for all three programs is collected in a coordinated process from hospitals and other medical facilities, including inpatient, outpatient, and standalone facilities. The data is collected or overseen by certified tumor registrars (CTRs) who are highly trained medical professionals to ensure complete and high quality data collection. The International Classification of Disease-Oncology (ICD-O) coding system is used to code topography (primary site) and morphology (histologic characteristics) of the collected cancers. Additional coding information is available in the NAACCR Data Standards & Data Dictionary (Volume II) available here: https://www.naaccr.org/data-standards-data-dictionary/.

Please note, the variables available in the CiNA Public Dataset are a subset of the full variable list collected. Many variables in the CiNA Public Dataset are aggregated and recoded for analysis.

Cancer Coding Changes Over Time

Several definitional changes occurred in some histology and behavior codes in ICD-O-3 that affected the inclusion and exclusion of reportable cancers diagnosed beginning in 2001.The changes predominately affected leukemias, lymphomas, and cancer of the ovary. One category of change between ICD-O-2 and ICD-O-3 is the manner in which leukemias and lymphomas are classified and coded. Although conversion of histology codes from ICD-O-2 to ICD-O-3 for cases diagnosed prior to 2001 helps minimize these differences, some minor differences may still exist, particularly with respect to some relatively rare lymphocytic cancers that can be coded to either leukemia or lymphoma.

Starting with ICD-O-3, several myelodysplastic diseases and syndromes are considered malignant, and, therefore, are now reportable for cases diagnosed in 2001 and later and are included in these data. Leukemias that represent a disease progression from one of the myelodysplastic diseases or syndromes diagnosed in 2001 and forward are no longer reportable.

For pediatric cancers, differences in incidence rates may be due to changes between the second and third edition of the International Classification of Childhood Cancers (ICCC). Two changes in the ICCC-3 classification are main contributors to this change. 1) Burkitt lymphoma and unspecified lymphoma, which were separated from non-Hodgkin lymphoma previously are combined with non-Hodgkin lymphoma; 2) Some lymphomas, which were grouped in the miscellaneous lymphoreticular neoplasms previously, are now included in the non-Hodgkin lymphoma category. Pilocytic astrocytoma is considered to have uncertain behavior in the published version of ICD-O-3, but is reportable as a malignant cancer in North America. Including the childhood astrocytomas in the category of malignant brain tumors may introduce differences between childhood brain cancer rates in North America compared to other areas of the world that may not include these tumors as malignant.

In addition, mesothelioma and Kaposi sarcoma cases are reported as separate categories. This change has little or no impact on most rates for specific cancers.

SEER*Stat Category	**Variable**	**Variable Type**
Age at Diagnosis	Age recode <1 year olds	Recode
Race, Sex, Year Dx, Registry, County	Year of Diagnosis	
	State/Province	
	Race/Ethnicity	Derived
	Country	Recode
	Race Recode (A, W, B, AIAN PRCDA, API, Othr)	Recode
	Race Recode (A, W, B, Othr, Unk)	Recode
	Sex	
Site and Morphology	Primary Site	
	Grade	
	Histology	
	Behavior	Recode
	Diagnostic confirmation	Recode
	Site recode ICD-O-3/WHO 2008	Derived
	ICCC site recode ICD-O-3/WHO 2008	Derived
Stage - LRD (Summary and Historic)	Summary Stage (SEER based on Diagnosis Year)	Recode
Extent of Disease - CS	Laterality	
Multiple Primary Fields	Record number recode	Recode
Census Tract attributes	Census Tract Poverty Indicator	
Other	Type of Reporting Source	Recode
ABSM	Yost Quintile (2000+, State Specific)	
County attributes 2010s	Metro/Non-Metro (2013 Beale)	Recode
County attributes 2000s	Metro/Non-Metro (2003 Beale)	Recode