Description of CiNA Public Use Dataset

The NAACCR CiNA-Public Dataset is a new CiNA product that was first available for 1995-2013 data. 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 after publication of the CiNA monographs; generally in the summer.

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 Use Agreement for access. There are other CiNA databases with more extensive variable set that require a proposal review, NAACCR IRB approval, and a "yes" consent by each participating registry. The CiNA Public Use the 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 comprised of two datasets. One allows a user to generate counts, rates and tends within the SEER*Stat system. This dataset includes age in the 19 age group categories. The second dataset allows the user to export the data as a case-listing to allow regression and other analysis in standard, statistical programs. This second dataset includes both age and site as categorical variables for US combined and Canada combined. A researcher will be able to access both data sets in the same SEER*Stat session but will only be able to export data from registries that have provided consent. SEER*Stat provides automatic cell suppression of <6 and scrambles the Patient IDs. *At this time, the database allowing exporting is still under development.*

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. A list of variables in the Public Use Dataset can be found here: https://www.naaccr.org/cina-public-use-data-set/

Data Use Agreement for Researcher Access to the Public Use Data Set
NAACCR CiNA-Public Dataset

These data are provided for the sole purpose of statistical reporting and analysis only. By using these data, you signify your agreement to comply with the following:

1. There will be no attempt to learn the identity of any person included in these data. If the identity of any person is discovered inadvertently, no disclosure or other use of the identity will be made, ***and I will notify NAACCR***.
   Initials required: _____
2. The identification or contact of individuals is prohibited. I will not discuss in any manner, with any unauthorized person, information that would lead to identification of individuals described in the Data furnished by NAACCR.
   Initials required: _____
3. I will not attempt either to link—***or permit others to link***—the data with individually identified records in another database.
   Initials required: _____
4. I will not either release—***or permit others to release***—the data—in full or in part—to any person. I will not share my password for data access with any other individuals. All members of a research team who have access to the data must sign this data-use agreement.
   Initials required: _____
5. I will not use—or permit others to use—the data in any way other than for statistical reporting and analysis for public health research purposes. I must notify NAACCR if I discover that there has been any other use of the data.
   Initials required: _____
6. I agree that all data provided shall remain the sole property of NAACCR and may not be copied or reproduced in any form or manner without NAACCR's prior written consent.
   Initials required: _____
7. I will cite the source of information in all publications. The appropriate citation is associated with the data file used.
   Initials required: _____
8. Uses of these data do not constitute an endorsement of the user's opinion or conclusions by NAACCR, or any central registry in US or Canada, and none should be inferred.
   Initials required: _____
9. I understand that calculating rates or other statistics based on small numbers can raise statistical issues concerning accuracy and usefulness. I will use appropriate caution when presenting and interpreting results based on less than 20 cases.
   Initials required: _____
10. I agree that any and all reports or analyses of the Data shall contain only aggregate data and no report of the Data containing statistical cells with less than six (6) subjects shall be released.
    Initials required: _____

My signature indicates that I agree to comply with the above stated provisions.

First Name: Last Name:
Organization: Email:
Phone:
Date: Signature:

Questions or issues, please contact Recinda Sherman, rsherman@naaccr.org.

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, 1995-2015 (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 – 2015. Confidentiality is maintained by aggregating data within individual records into categories, the number of which depends on whether analysis is run on individual states/registries or nation/North America.

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). Please note, counts less than six may be released for other variables including behavior, stage or histology. However, these variables are not considered identifying variables.
- 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. Current CiNA Public Use dataset contains 46 states plus DC and PR for the US, and 12 Canadian provinces covering 93% of the US and 64% of the Canada population.

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.