Information for CiNA Data Set Investigators

1. Eligibility for data access and description of data file contents

- You must be a member of the North American Association of Central Cancer Registries (NAACCR), or an employee of a member organization (http://www.naaccr.org/Membership/MembershipDirectory.aspx), to receive access to CiNA Deluxe data files. [Information on NAACCR membership can be found at http://www.naaccr.org/Membership/BecomeaMember.aspx]

- If you are not a NAACCR member, contact a NAACCR registry member, a member organization, or Recinda Sherman, NAACCR Program Manager of Data Use and Research (rsherman@naaccr.org), to discuss your proposal idea to determine if it is feasible based on available CiNA Deluxe data. **At least one co-investigator for all CiNA Data Sets applications must be a NAACCR member**. Researchers who are not NAACCR members may serve as co-investigators and/or apply for individual membership.

- The NAACCR Data Sets are provided to researchers within the SEER*Stat statistical software program. You must be familiar with SEER*Stat (http://seer.cancer.gov/seerstat/index.html) and cancer statistics to ensure accurate data analysis.

2. CiNA Data Sets

2.1. Source of data

Data in the CiNA Data Sets are from the NAACCR Call for Data. Population-based central cancer registries in the United States and Canada submit cancer incidence data, from 1995 to the most current year, to NAACCR in December each year for its Call for Data.

2.2. Data quality and inclusion criteria

CiNA Data Sets include data only from registries that meet all of the CiNA Monograph inclusion criteria (see the table below) for at least three consecutive years. Data quality and completeness vary by registry and year of diagnosis. You are advised to assess the quality and completeness of the data before proceeding with data analysis. Researchers may have to limit data analysis to a subset of registries or to certain diagnosis years because of missing or low quality data.

Criteria	CiNA Deluxe Inclusion Requirement
Completeness	≥90%
% Passing EDITS	100%
Death-Certificate-Only Cases	<=5%
Timeliness	Received by December due date (within 23 months of the close of the diagnosis year)
Duplicate Reports	<=2/1,000 cases
Missing Sex, Age, County	<=3%
Missing Race	<=5%

2.3. Timeliness

Registries are required to submit data for the NAACCR Call for Data within 23 months of the close of a diagnosis year (i.e., December 1, 2013, for all cases diagnosed between 1995 and 2011). The CiNA Data Sets are available for use about 30 months after the close of the diagnosis year (i.e., July 2014 for all cases diagnosed between 1995 and 2011).

2.4. Data elements

NAACCR collects data on a limited set of variables for certification, publication and evaluation. Three types of files are available for researchers:

- **CiNA Public Use**
- **CiNA Research Dataset, "standar"d with aggregated age-groups or "special" with with single years of age**
- **CiNA Survival/Prevalence**

A variable list is available here: https://www.naaccr.org/cina-data-products-overview/

You are advised to review the variable list to ensure that data elements that you intend to use are available in the CiNA Data Sets. Further information regarding values and codes in the CiNA Data Sets may be found in the NAACCR Data Standards and Data Dictionary, Volume II
https://www.naaccr.org/data-standards-data-dictionary/ .
If you still have questions after reviewing these materials, please send an email to Recinda Sherman rsherman@naaccr.org.

3. Application requesting access to CiNA Deluxe data files.

To request data, researchers access the Data Request Tracking (DaRT) System

accessible here: https://www.naaccr.org/cina-data-products-overview/

The application consists of:

- Study proposal using the CiNA Deluxe proposal template
- Biographical sketches for all investigators
- Data Confidentiality Agreement for NAACCR Researchers Form signed for every person who will access the data

4. Application review process

4.1. Review process

- Complete applications are distributed to the NAACCR Research Application Review (RApR) Workgroup by the chairperson, who designates two reviewers (primary and secondary). An employee of Information Management System, Inc. (IMS), the company that houses the CiNA Deluxe data, also reviews the technical aspects of the proposal. Reviewers will have a minimum of two weeks to complete the application review.

- The reviewers present their review during the next RApR Workgroup monthly conference call (the second Thursday).

- Based on reviewers' comments and the committee's discussion, the RApR Workgroup will make one of three decisions: Approved, Approved Pending Revisions Requested, or Rejected.

- The primary and secondary reviewers finalize their reviews and submit written comments through the DaRT System approximately one week after the RApR Workgroup conference call**.**

- The NAACCR Program Manager of Data Use & Research (Recinda Sherman) notifies the Principal Investigator of the application about the decision, and provides relevant reviewers' comments, via DaRTS.

4.2. Review criteria

The RApR Workgroup examines the proposal for:

- Experience of investigators in analyzing cancer surveillance data
- Clarity of research questions and objectives
- Availability of data items to address stated research questions
- Appropriateness of analytic approach

4.3. NAACCR consent process and continuing monitoring

- You are required to complete and submit annual DUAs to maintain access to the CiNA data

- You are required to send pre-submission manuscripts for NAACCR Review prior to publication

- NAACCR will contact all registries whose data are eligible for your project to request consent for the use of their data in the study. Registries are usually given a minimum of two weeks to respond.

- **Note:** NAACCR requires active consent when the proposed project requests area identifiers (e.g., county code), single-years of age, linkage with other datasets, or any new or unique methodology. For all other projects, consent is passive.

- If a registry does not respond to a request, NAACCR assumes the registry consents to projects requiring passive consent and does not consent to projects requiring active consent.

4.4. Length of the application process

- Approximately a month - RApR Workgroup review process
- Approximately 3 weeks - registry consent process
- Approximately 2 weeks - IMS prepares the data files in SEER*Stat

You should be able to gain access to the data in SEER*Stat in approximately 3 months. ***Reminder: The workgroup meets on the second Thursday of every month.***

4.5. Access to CiNA Deluxe data through SEER*Stat

- You will receive an email notification from Information Management Services, Inc. (IMS) when access to your NAACCR CiNA Data Set is available.

- The data files will contain data from consenting registries only.

5. Confidentiality and responsibilities

- The researche must formally agree and acknowledge that patient confidentiality is of the utmost importance in using the data and in presenting or publishing research results.

- Your agreement with NAACCR states specifically that the data must be used only for the purpose defined in the approved proposal. Any other use violates this agreement.

- Access to the data file will be terminated at the conclusion of the project or if a researcher does not provide an annual DUA.

6. Publication and review instructions

Upon completion of the study, NAACCR reserves the right to review findings and ensure compliance with privacy conditions.

6.1 Data quality issues

- It is standard practice for researchers to assess the quality of data variables used in their analyses. If, during the course of the research, an investigator identifies any unanticipated data quality issues, the investigator shall provide a brief written summary of those issues to Recinda Sherman (rsherman@naaccr.org) at NAACCR.

6.3. Approval of NAACCR Scientific Editorial Board

- Any manuscript resulting from the research project must be submitted to the NAACCR Scientific Editorial Board (rsheramn@naaccr.org or through DaRT) for review and approval before being submitted for publication. The review process takes a minimum of two weeks.

6.4. Acknowledgment

Please acknowledge NAACCR and the use of its CiNA Data Sets files in publications or presentations for projects using the CINA data. Suggested citations are available for each unique data set in SEER*Stat.

6.5. Courtesy manuscript copy

Once a manuscript has been accepted for publication, investigators should send a courtesy copy to Recinda Sherman (rsherman@naaccr.org) at NAACCR for via DaRT.

7. Special notes regarding CiNA Deluxe data

CiNA Research data are complex and sometimes difficult to analyze. Please see the following comments about selected data items.

7.1. Race and Hispanic Ethnicity Variables

- **NAACCR CiNA Analytic File 1995-20XX, for Expanded Races**. Researchers should use this file for all analyses that include race as an analytic variable. This file should be the default research file since it includes the maximum number of registries with data available for studies.

- **NAACCR CiNA Analytic File 1995-20XX, for NHIAv2 Origin:** This file is the source for all analyses that include Hispanic ethnicity as an analytic variable. The NAACCR Hispanic/Latinx Identification Algorithm, version 2 (NHIA v2) is a method used by all but 2 cancer registries in the United States to enhance identification of persons of Latinx heritage diagnosed with cancer. Information describing the NHIA algorithm (which is now run through the Hispanic and Asian/Pacific Islander Identification Algorithm (NHAPIA) is available on the NAACCR website (http://www.naaccr.org/Research/DataAnalysisTools.aspx). All investigators should be

cautious in their interpretation of these data because in areas of small Latinx populations, methods that indirectly identify Latinx, such as NHIA, can overestimate the ethnicity-specific counts of cancer cases. Also, even small errors in population estimates of Latinos can affect the magnitude of cancer rates.

In the NAACCR NHIA Analytic file, race is combined with ethnicity to produce the following population categories: Hispanic (all races combined), non-Hispanic white, and non-Hispanic black.

7.2. County Information

- Some states are prohibited from submitting county identifiers to NAACCR. The standard CINA Deluxe file contains a rural-urban indicator rather than county identifiers.

- In 2002, Colorado created a new county, Broomfield County. Thus, the 2000 and 2001Census population estimates did not contain data for Broomfield County. Population characteristics for Broomfield County were created for 2002 and later in order to produce accurate county-specific population counts by race, Hispanic origin, age, and sex. These changes do not affect analyses of state-level data. Similarly, in 2007 and 2008, Alaska created five new counties and deleted three and the new boundaries are used for 2011 data forward. Researchers must be familiar with their subject area in order to select the correct counties over time.

7.3. Urban Rural Continuum

The Urban Rural Continuum variables (Beale codes) are used as measures of urbanization on a county level. These variables make use of information from the preceding census (i.e., 2003 codes are based on the 2000 census). The 1993, 2003 and 2013 County-Level Urban Rural Continuum codes are available in the file. Please see the following website for more information on the codes and their definition: http://www.ers.usda.gov/data-products/rural-urban-continuum-codes

7.4. Impact of Hurricane Katrina/Maria

Hurricane Katrina severely affected cancer surveillance in 2005 in Alabama, Louisiana, Mississippi, and Texas. Case counts in CINA Deluxe for these states are for the entire year. However, age-adjusted rates are calculated only on the number of cases diagnosed between January 1, 2005, and June 30, 2005. To include case data for the entire year for

these 4 states, the box in the SEER*State query definition window "Cases in Research Database" located within the "Selection" tab must be unchecked. Similarly, Hurricane Maria severely impacted data collection in Puerto Rico for the 2017 data.

7.5. Duplication of Incidence Cases in CiNA Deluxe

The CiNA Deluxe file contains data from all registries, including those that are population-based but not state-based (e.g., the metropolitan areas of Detroit, Greater Bay Area, Los Angeles, Greater California, and Seattle). For the states encompassing these metropolitan areas, cases in the file are duplicated. To avoid double counting, researchers must exclude overlapping metropolitan areas when running analyses that are not stratified by registry.

7.6. Analyses involving time trends (caution)

It is very important that any analyses involving time trends include only data from registries that have data for all years in the time interval under investigation. It has been demonstrated that registries that do not meet this criterion may have rates and trends that differ from registries for which data are complete. Including partial data from such states can skew overall trend estimates, lead to biases in the data, and result in inaccurate conclusions.

7.7. Stage of disease at diagnosis

- **Summary Stage 1977 and SEER Summary Stage 2000 (U.S. Only)**

NAACCR has evaluated the effect of coding changes in Summary Stage 2000 compared with Summary Stage 1977. SEER Summary Stage 1977 is applied to the 1995 to 2000 cases, and SEER Summary Stage 2000 is applied to the 2001-2003 cases. This report is available on NAACCR web site:
http://www.naaccr.org/DataandPublications/DataQuality.aspx

- **Collaborative Stage Derived Summary Stage 2000 (2004 and forward)**

The Derived Summary Stage 2000 is the coding schema used for all cases diagnosed 2004 and after. It is calculated using data items from the Collaborative Stage System. NAACCR has evaluated the comparability of the Derived Summary Stage 2000 and the Directly Coded Summary Stage 2000 using data from cancer registries. The report is available on http://www.ncbi.nlm.nih.gov/pubmed/21688742

7.8. Using dates of treatment

Two variables are available as standards for the date of first treatment, one used by the SEER program and the other by the Commission on Cancer (CoC). Both include the year and month of first treatment. Registries that are part of the SEER program are required to use the SEER data collection standard. Registries in the NPCR program have the option to use either the SEER or CoC standard. Thus, both variables need to be used with caution since the two standard setters have applied different meanings to the variable. Investigators using these variables must refer to *NAACCR Standards for Cancer Registries, Volume II: Data Standards and Data Dictionary*. This document can be found on NAACCR web site at: https://www.naaccr.org/data-standards-data-dictionary/

7.9. For Canadian Registry Research

- Canadian registries do not collect information on race.

- The Rural Urban Continuum codes are not available for Canada.

- Canadian Registries have implemented Collaborative Stage Derived Summary Stage 2000, for cases diagnosed in 2004 and later. The data are available in the file as submitted by provinces; however, these data are incomplete. Summary Stage 1977 and SEER Summary Stage 2000 are not available.

- Canadian registries do not follow NAACCR standards for type of reporting source, and this data element is not standard among Canadian registries.