

## **Protocol for Assessing Duplicate Cases for CiNA (1995-2018), Registry Certification (2018), and 12 Month Data (2019)**

This protocol will be used to evaluate the number of duplicate records that reside on your registry database and have not been identified or corrected using regular matching, linkages, or other registry protocols. This assessment of duplicates must be completed in order for the data from your registry to be eligible for CiNA research files (1995-2018), Registry Certification (2018), and evaluation of your 12 Month Data (2019).

**PLEASE NOTE:** The protocol must be completed separately for the CiNA and Certification Submission files, as well as for your 12 month Data Submission.

Completing the protocol is **mandatory** for CiNA and Certification submission  
and **recommended** for 12 Month Data.

### **Duplicate Assessment Protocol for CiNA:**

1. Extract a list of **all** records of eligible primary tumors for 1995-2017 (include cases obtained through data exchange agreements with other central cancer registries, federal facilities like the VA, and other non-hospital data sources). If your registry did not start on or before 1995, then select a list from all available years of complete data.
2. Select one or more geographic areas that will produce a list of cases that meets the *sample size requirement* stated in the table below. If the total number of records in your registry's database is smaller, conduct the protocol for assessing duplicate cancer records on your entire database.

### **Duplicate Assessment Protocol for Registry Certification:**

1. Follow the Protocol for CiNA Submission, except perform the protocol on 2018 cases only rather than 1995-2017 cases.
2. Select one or more geographic areas that will produce a list of cases that meets the sample size requirement stated in the table below. If the total number of records in your registry's database is smaller than the required sample size, conduct the protocol for assessing duplicate cancer records on your entire database.

### **Duplicate Assessment Protocol for 12 Month Data:**

Follow the Protocol for Registry Certification, except perform the protocol on 2019 cases rather than 2018 cases.

\*\*\*\*\*  
**SAMPLE SIZE REQUIREMENTS**

- Any registry with fewer than 4,500 cases for the time period **MUST** use all available cases to conduct the duplicate report assessment protocol.
- Sample size designation for the protocol is based upon unresolved duplication rate reported by each registry for the 1995-2017 CiNA data file submission.
  - The minimum sample size of 4,500 is required for registries reporting unresolved duplicates of <0.50 per 1,000.
  - A 15,000 sample is required of registries with reported duplicates of 0.5 to <2.0 per thousand *or* did not submit 2012-2016 data for the 2019 CiNA publications.

- A 20,000 sample is required of registries with reported duplicates of 2.0 to <3.0 per thousand.
- Registries over 3.0 per thousand are required to use all cases.

**Use the table below to identify the sample size for your registry.**

<b>Sample Size</b>	<b>Registry</b>
4,500	All registries not listed below should use a sample size of 4,500. If the entire registry submission is smaller than 4,500, the registry should use all their cases.
15,000	Maine, Massachusetts, Minnesota, Pennsylvania, South Dakota, Wisconsin, Quebec, Yukon
20,000	none

**DATA ITEMS USED TO DETERMINE DUPLICATES**

1. Data items in each record that can be helpful in determining a duplicate report are:
  - a. last name
  - b. first name
  - c. social security number (U.S.) or health insurance number (Canada)
  - d. date of birth
  - e. SEER cancer site group
  - f. middle name
  - g. age at diagnosis
  - h. sex
  - i. race
  - j. date of diagnosis
  - k. tumor sequence number
  - l. primary site code
  - m. morphology code
  - n. laterality
  - o. stage of disease
  
2. List all reports of eligible primary tumors selected for the protocol. Sort the list by:
  - a. last name
  - b. first name
  - c. social security number or health insurance number
  - d. date of birth
  - e. SEER cancer site group

It may be useful to include the patient identification number. Also, add any other variables to the list that you feel would assist you in resolving the identities of duplicate records.

Determine whether duplicate last names and first names or meaningful equivalent names (e.g. Bill and William) reflect reports of multiple primaries, using the SEER Rules for Determining Multiple Primary Cancers.

- Using the following table, as a guideline, review the list to identify potential duplicate records of the same tumor:

MATCH CRITERIA IN IDENTIFYING POTENTIAL MATCHES							
Last Name	First Name	Sex	D.O.B	SSN	Primary Site	Morphology	Date of Dx
x		x	x				
x	x	x		x			
	X		x		x		
		x	x	x			
x	x		x				
x	x			x			
x	x		x	x			
x	x	x	x	x			
x			x	x			
x	x		x		x		
x		x		x		x	x
x	x		x	x		x	x
	x	x			x		
x	x				x		x
x	x		x		x	x	x
x	x			x	x	x	x

A duplicate record is considered to meet the criteria of any of the following:

- All elements in the table are identical. This is a duplicate record and should be counted as a duplicate record on the Identification of Duplicates Form.
- Probable match (one or more facilities reporting similar data) most elements are the same and a few are so similar they have a great potential to be the same case report.

An example would be same last name, meaningful equivalent first name (Virginia and Ginny), same diagnosis date, and same birth month and year but: different day of birth (11/12/49 and 11/21/49) or two digits of SSN different (999-99-9901 and 999 99-9910) or similar site codes (C161 and C169) or similar morphology codes (8140/3 and 8490/3). The registry should determine, with assistance if necessary from the reporting facilities, whether the two records are duplicate records.

- Report the number of duplicates found and the exact sample size on the Duplicate Protocol Form on the NAACCR Call for Data submission website.
- Resolve all duplicate records before submission to NAACCR.