



June 15-17, 2021



Cancer Data
Registry of Idaho

Calculating Cancer Incidence using the new NCI Census Tract Populations

Christopher Johnson, MPH

Cancer Data Registry of Idaho, USA

Coauthors

- Steve Scoppa, IMS
- Mandi Yu, NCI
- Martin Holdrich, Woods & Poole
- Todd Gibson, IMS
- Recinda Sherman, NAACCR



Acknowledgments and Disclaimer

- This project was funded in whole with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201800006I and the Centers for Disease Control and Prevention, Department of Health and Human Services, under Cooperative Agreement 1NU58DP006270, and to the Cancer Data Registry of Idaho, Idaho Hospital Association.
- The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention or the National Cancer Institute.

Objectives

1. Participants will learn how to embed the new NCI Census Tract Populations in local SEER*Stat databases
2. Participants will learn how to calculate cancer incidence statistics for CDC Sub-County Geographies and NCI/NAACCR Cancer Reporting Zones



Objectives

1. Participants will learn how to embed the new NCI Census Tract Populations in local SEER*Stat databases
2. Participants will learn how to calculate cancer incidence statistics for CDC Sub-County Geographies and NCI/NAACCR Cancer Reporting Zones



Step 1 – Prep the Case Data

NAACCR XML Utility

XML to Flat

- i. Read Nov 2020 Data Submission File
(NPCR or SEER, not NAACCR)
XML V18
- ii. Write Flat V18

NAACCR XML Utility v7.9

File Help

XML to Flat: transform a given NAACCR XML file into the corresponding Flat file (fixed-columns)

Source XML File: E:\CANCER\Call_For_Data\2020 Nov Call for Data\NPCR\DataSubmission\ID9519V18.xml Browse...

Source File format: Compressed NAACCR 18 Incidence Num lines: <?> Num patients: <?> Num tumors: <?> File size: 33.8 MB

Please review and/or change the following options. Once you are ready, click the process button at the bottom of the page.

Target Flat File: R:\EPI\CANCER\NAACCR\Conferences\conf2021\CI_presentation\ID9519V18.txt Browse... Compression: None

Processing Options

- ☐ When reading the items, validate their value.
If this option is checked, each value will be validated against the item's data type defined in the dictionary.
- ☐ When reading the file, ignore unknown items.
If this option is checked, unknown items will be ignored. Otherwise a warning will be reported.
- ☒ When writing the items, apply zero-padding rules (this might change the actual values being written).
If this option is checked, items defining a padding rule (usually left 0-padded) will have their value modified to honor the definition.
- ☐ When writing the items, report an error if a value is too long.
If this option is checked, a warning will be reported if an item in the XML has a value that is too long for the flat file. Otherwise the value will be silently cut-off to the maximum length allowed. Note that this option doesn't affect items that allow unlimited-text; those will be silently cut-off regardless.
- ☒ When reading the file, automatically translate renamed NAACCR XML IDs.
If this option is checked, the application will automatically translate the IDs that were shortened in NAACCR 18. You can use this option to read an XML file with the old IDs and re-create it with the new ones.
- ☒ When reading the file, use strict rules for the XML namespaces.
If this option is checked, non-standard XML tags/attributes will need to be defined in a proper namespace using a namespace prefix.

Exclude/Include following items: Comma-separated list of items (NAACCR ID or NAACCR Number) to exclude/include when reading or writing the file. ☒ Exclusion list ☐ Inclusion list

All the user-defined dictionaries have been provided

Process Source File

Step 2 – Procure the Population Data

Download U.S. Population Data – 1969–2019

Released February 2021

Datasets

U.S. Mortality Data

U.S. Population Data –

Download U.S. Population Data

Single Year of Age County Population Estimates

Estimates Used in SEER*Stat Software

Standard Population Data +

County Attributes +

Expected Survival Life Tables +

SEER Linked Databases

Files containing U.S. population data are provided for analysis using your own statistical software or using [SEER*Stat](#) (you must use the [SEER*Prep](#) software to create a database for use in SEER*Stat).

Notes

- The U.S. Census Bureau annually releases [unbridged population estimates for five-year age groups and race at the county level](#). The Census Bureau does not release bridged race estimates by single year of age at the county level due to concerns about the reliability of these estimates. However, these estimates are provided to the National Cancer Institute to meet programmatic needs such as the creation of age groupings that differ from the standard groupings used by the Census Bureau. Users of the single-year-of-age county-level bridged race population estimates should carefully consider the limited reliability of these estimates.
- Annual data releases include both an additional year of data and updates to previous estimates for all years going back to the last decennial census (see [Population Estimates Used in NCI's SEER*Stat Software](#)).

Available Files

Populations are available for the following time periods and "races":

- 1969-2019 County-level: White, Black, Other;
- 1990-2019 County-level: Expanded Races (White, Black, American Indian/Alaska Native, Asian/Pacific Islander) by Origin



Step 3 – Census Tract Attributes



Home

Cancer Statistics ▾

SEER Data & Software ▾

Registry Operations ▾

News

About

Home ▸ SEER Data & Software ▸ U.S. Datasets ▸ County Attributes ▸ Time-dependent County Attributes

Time-dependent County Attributes

Datasets

U.S. Mortality Data

U.S. Population Data +

Standard Population Data +

County Attributes -

Static County Attributes

Time-dependent County Attributes

Expected Survival Life Tables +

SEER Linked Databases

The time-dependent county attributes include a socioeconomic status (SES) index and seven county-level SES attributes that were used to construct the index. The attributes and the index are estimated at various time points using data obtained from the 1990 and 2000 U.S. Decennial Census long form survey, and a series of American Community Survey (ACS) 5-year estimates from 2006 to 2019. At each time point, a SES index is constructed from these attributes using a factor analysis (Yu et al. 2014). The attributes and the index are then linked to cancer cases/deaths at the county level by matching the survey data year and diagnosis/death year. Thus, for any of the seven attributes and index, cancer cases/deaths in the same county may have different values for different years. The mapping of the SES attributes/index and diagnosis or death year are shown in Table 1.

Table 1.

Years	Data Source	Geographic Definition	Counties
1990-1993	1990 Census	1990	All counties
1994-1995	1990 Census	1990	All counties except combined Aleutians East Borough and Aleutians West Borough, Alaska, combined Dillingham Census Area and Lake and Peninsula Borough, Alaska, combined Skagway-Hoonan-Angoon Census Area and Yakutat City and Borough, Alaska and combined La Paz County and Yuma County, Arizona
			Aleutians East Borough, Alaska (02013) Aleutians West Borough, Alaska (20016) Dillingham Census Area, Alaska (02070) Lake and Peninsula Borough, Alaska (02012)



Step 3 – Census Tract Attributes



1. Census Tract 11-digit FIPS Code
2. NCI/NAACCR Cancer Reporting Zones
3. CDC Sub-County 5k and 20k Geography
4. Yost Scores (and 7 components)
5. NAACCR Poverty
6. RUCA Codes, URIC Codes
7. Other area-based measures...
8. Year (for time dependent attributes)



Step 3 – Census Tract Attributes



1. Download the Census Tract Attributes .txd (data) file and the .dd (dictionary) file from NCI website.
2. Use SEER*Prep to build the Census Tract Attributes database locally.
3. Details to follow...



Step 4 – SEER*Prep DD for Census Tract Incidence

 **NATIONAL CANCER INSTITUTE**
Surveillance, Epidemiology, and End Results Program

[f](#) [t](#) [e](#) [r](#) [+](#)

Search SEER 

HomeCancer Statistics ▼SEER Data & Software ▼Registry Operations ▼NewsAbout

[Home](#) ▶ [SEER Data & Software](#) ▶ [Statistical Software](#) ▶ [SEER*Prep Software](#)

SEER*Prep Software

Version 2.6.0 and Database Description Files released March 10, 2021

Statistical Software

SEER*Stat +

SEER*Prep -

Input File Formats

Utility Programs

Getting Help

FAQs

How to use SEER*Prep

HD*Calc +

The SEER*Prep software converts ASCII text data files to the SEER*Stat database format, allowing you to analyze your cancer data using SEER*Stat. SEER*Prep performs two main functions: it converts text data to the specific binary format required by SEER*Stat, and it creates the SEER*Stat data dictionary. SEER*Stat 8.3.9 or later is required to analyze any database generated with this version of SEER*Prep. If you are a new SEER*Prep user, see [How to use SEER*Prep to Create a Database](#).

[Readme.txt](#) (TXT, 23 KB) distributed with the software includes a description of the revisions made for this release.

[SEER Incidence Site Recode](#) and [SEER Cause of Death Recode](#) describe current and historical SEER*Prep algorithms related to these variables.

Download SEER*Prep

If you have problems downloading SEER*Prep, please send an email to seerprep@imsweb.com.

Supported File Formats

SEER*Prep can be used to convert incidence, mortality, expected survival, and standard population data stored in one of the formats listed below. The [Input File Formats](#) section contains more specific information about these file formats and general rules that must be followed when creating input data files.

Incidence:

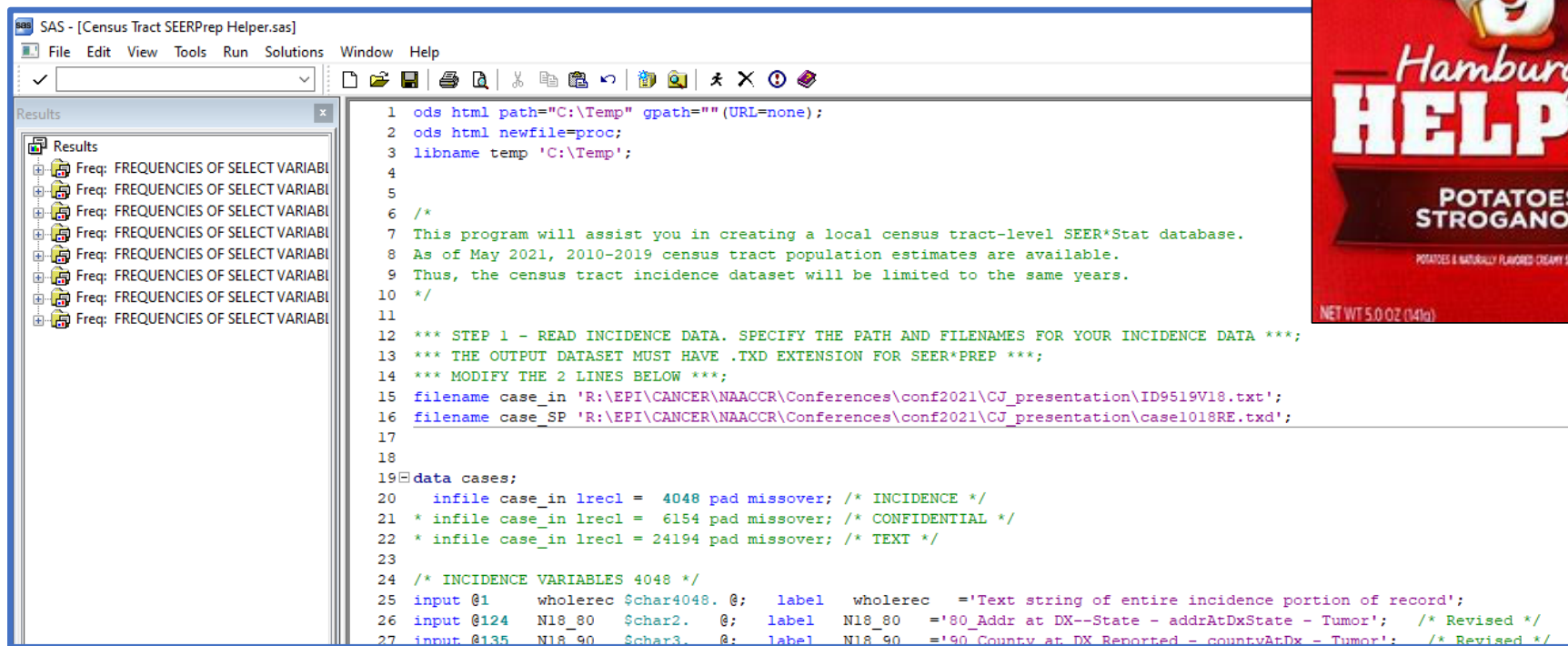
- NAACCR - 4048 Byte Format
- Global - 234 Byte Format (For the difference between the NAACCR and the Global formats [see the FAQs.](#))



forum.org



Step 5 – SAS Helper



```
1 ods html path="C:\Temp" gpath="" (URL=none);
2 ods html newfile=proc;
3 libname temp 'C:\Temp';
4
5
6 /*
7 This program will assist you in creating a local census tract-level SEER*Stat database.
8 As of May 2021, 2010-2019 census tract population estimates are available.
9 Thus, the census tract incidence dataset will be limited to the same years.
10 */
11
12 *** STEP 1 - READ INCIDENCE DATA. SPECIFY THE PATH AND FILENAMES FOR YOUR INCIDENCE DATA ***;
13 *** THE OUTPUT DATASET MUST HAVE .TXD EXTENSION FOR SEER*PREP ***;
14 *** MODIFY THE 2 LINES BELOW ***;
15 filename case_in 'R:\EPI\CANCER\NAACCR\Conferences\conf2021\CJ_presentation\ID9519V18.txt';
16 filename case_SP 'R:\EPI\CANCER\NAACCR\Conferences\conf2021\CJ_presentation\case1018RE.txd';
17
18
19 data cases;
20   infile case_in lrecl = 4048 pad missover; /* INCIDENCE */
21   * infile case_in lrecl = 6154 pad missover; /* CONFIDENTIAL */
22   * infile case_in lrecl = 24194 pad missover; /* TEXT */
23
24 /* INCIDENCE VARIABLES 4048 */
25 input @1 wholerec $char4048. @; label wholerec = 'Text string of entire incidence portion of record';
26 input @124 N18_80 $char2. @; label N18_80 = '80_Addr at DX--State - addrAtDxState - Tumor'; /* Revised */
27 input @135 N18_90 $char3. @; label N18_90 = '90_County at DX Reported - countyAtDx - Tumor'; /* Revised */
```



Step 6 – SEER*Prep

SEER*Prep 2.6.0 - [r:\epi\cancer\naacr\conferences\conf2021\cj_presentation\naacr4048.d03042021_censustract.dd]

File Execute Help

☐ Summary Files

Database Name: Idaho 2010-2018 NCI Census Tract 2010 Geo

Input File(s): Record Length: 4048
R:\EPI\CANCER\NAACR\Conferences\conf2021\cj_presentation\case1018RE.txd

Add... Remove Up Dn

Input Population File(s): Record Length: 38
R:\EPI\CANCER\NAACR\Conferences\conf2021\cj_presentation\pops1018RE.txd

Add... Remove Up Dn

When Linking to Populations Use:

Age: Age recode with <1 year olds

Default Std: 2000 US Std Population (19 age groups - Census P25-1130)

Race: Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)

Population Count Location: Edit...
Start Col: 30, Length: 9

Study Cutoff Date for Survival: Edit...
Calculation: December, 2018

Citation/Message/Footnote...

Variable	Case Start Col	Pop ...	Length	Re...	Index
Year of diagnosis	544	1	4	Yes	Yes
Addr at DX--state	124	5	2	Yes	Yes
State-county	3600	7	5	Yes	Yes
State	3600	7	2	Yes	Yes
PRCDA 2017	3600	7	5	Yes	Yes
PRCDA Region	3600	7	2	Yes	Yes
County at DX	135	9	3	Yes	Yes
County at DX Analysis	150	9	3	Yes	Yes
County	3602	9	3	Yes	Yes
* Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)	N/A	14	1	Yes	Yes
Sex	222	16	1	Yes	Yes
* Age recode with <1 year olds	N/A	17	2		Yes
Census Tract Prefix (first 4 digits)	3925	19	4	Yes	Yes
Census Tract Suffix (last 2 digits)	3929	23	2	Yes	Yes
Primary Site - labeled	555		3	Yes	Yes
Primary Site	555		3	Yes	Yes
Histologic Type ICD-O-3	564		4	Yes	Yes
Behavior code ICD-O-3	568		1	Yes	Yes
Grade	569		1	Yes	Yes
Grade path value	570		1	Yes	Yes
Grade path system	571		1	Yes	Yes
Grade Clinical	1286		1	Yes	Yes
Grade Pathological	1287		1	Yes	Yes
Grade Post Therapy	1288		1	Yes	Yes
Diagnostic Confirmation	576		1	Yes	Yes
Ambiguous Terminology DX	580		1	Yes	Yes

* - Derived Variable (Case File Only)

Add... Edit... Remove

Step 6 – SEER*Prep

SEER*Prep Create Database Report

Database:

Idaho 2010-2018 NCI Census Tract 2010 Geo

When linking to populations use:

Age: Age recode with <1 year olds

Default standard population: 2000 US Std Population (19 age groups - Census P25-1130)

Race: Race and origin recode (NHW, NHB, NHAIAN, NHAPI, Hispanic)

Citation:

SEER*Stat Database: Idaho Census Tract using NCI Pops 2010 Geo . Created on May 31, 2021.

Message:

Not specified

User-Defined Data Location:

C:\data\cancer\seerstat\

Study cutoff date:

December, 2018



Step 7 – Modify the database.ini File

```
Name=Idaho 2010-2018 NCI Census Tract 2010 Geo  
Citation=SEER*Stat Database: Idaho Census Tract using NCI Pops 2010 Geo . Created on May 31, 2021.  
ID=1622499118  
LinkDBID=1622553293  
LinkFields=State-county~Census Tract Prefix (first 4 digits)~Census Tract Suffix (last 2 digits)
```

After the ID line, comment out these lines by removing semicolons.

LinkDBID tells SEER*Stat to link one SEER*Stat database to another. The value of the LinkDBID will be different from what is shown – it will be the database ID from your local census tract attributes file.

LinkFields should be exactly like shown.



Objectives

1. Participants will learn how to embed the new NCI Census Tract Populations in local SEER*Stat databases
2. Participants will learn how to calculate cancer incidence statistics for CDC Sub-County Geographies and NCI/NAACCR Cancer Reporting Zones



Analysis by Census Tract in SEER*Stat

Rate Session-1

Data | Statistic | Selection | Table | Output

ID	Link ID	Database Name
1535403108	1432844115	Idaho Mortality 1970-2017
1540768560	1432844115	Idaho Cases 1970-2017 (Prelim Annual Report)
1537656481	1432844115	Idaho Mortality 2010-2017 Geocoded Census Tract
1544482367	1432844115	Idaho Cases 2007-2017 Census Tract 2010 - Differential Tract Growth ACS 5-yr series log linear r
1544482561	1432844115	Idaho Cases 2007-2017 Census Tract 2010
1547484515	1432844115	Idaho Cases 2006-2017 Census Tract 2010 - Differential Tract Growth ACS 5-yr series log linear r
1549399599	1432844115	Idaho Mortality 2010-2017 Geocoded Census Tract - Differential Tract Growth ACS 5-yr series log
1555789052	1432844115	Idaho Trauma Registry 2015-2016 Census Tract 2010 - Differential Tract Growth ACS 5-yr series
1565988461		Idaho Cases 1970-2017 (Prelim Annual Report) POST SSA - reported alive survival
1568128668		Idaho Peds Cases 1990-2005 and All Cases 2006-2017 Census Tract 2010
1571505042		Idaho Mortality 1970-2018
1572115434		Idaho Cases 1970-2018 (Prelim Annual Report)
1574546101		Idaho Cases 2000-2016 PRE SSA
1574546914		Idaho Cases 2000-2016 POST SSA
1574723501	1578672991	Idaho Cases 1990-2018 Census Tract 2010 - Differential Tract Growth
1574278645	1578672991	Idaho Mortality 2010-2018 Geocoded Census Tract - Differential Tract Growth ACS 5-yr series log
1596745899		Idaho Cases 1970-2018 (Prelim Annual Report - August 2020)
1603828093		Idaho Mortality 1970-2019
1603753795		Idaho Cases 1970-2019 (2018 Annual Report - October 2020)
1604691101		Idaho Cases 1990-2019 Census Tract 2010 - Differential Tract Growth
1609179369	1432844115	Idaho Mortality 2010-2019 Geocoded Census Tract - Differential Tract Growth ACS 5-yr series log
1615307188		Idaho Cases 2019 (2019 Prelim Annual Report - March 2021)
1622499118	1622553293	Idaho 2010-2018 NCI Census Tract 2010 Geo

Find | Original Sort Order | * Change Linked Database...

Age Variable: Age recode with <1 year olds

Suggested citation for the selected database:
SEER*Stat Database: Idaho Census Tract using NCI Pops 2010 Geo . Created on May 31, 2021.

Analysis by Census Tract in SEER*Stat

Rate Session-1

Data | Statistic | **Selection** | Table | Output

Select Only

☐ Malignant Behavior ☒ Known Age [View Statements](#)

☒ Male or Female Sex ☐ Cases in Research Database

Age at Diagnosis (Std Pop, Pop, Case Files)

Edit...
Copy
Clear

Race, Sex, Year Dx, Registry, County (Pop, Case Files)

{Race, Sex, Year Dx, Registry, County, Year of diagnosis} = '2010','2011','2012','2013','2014','2015','2016','2017','2018'

Edit...
Copy
Clear

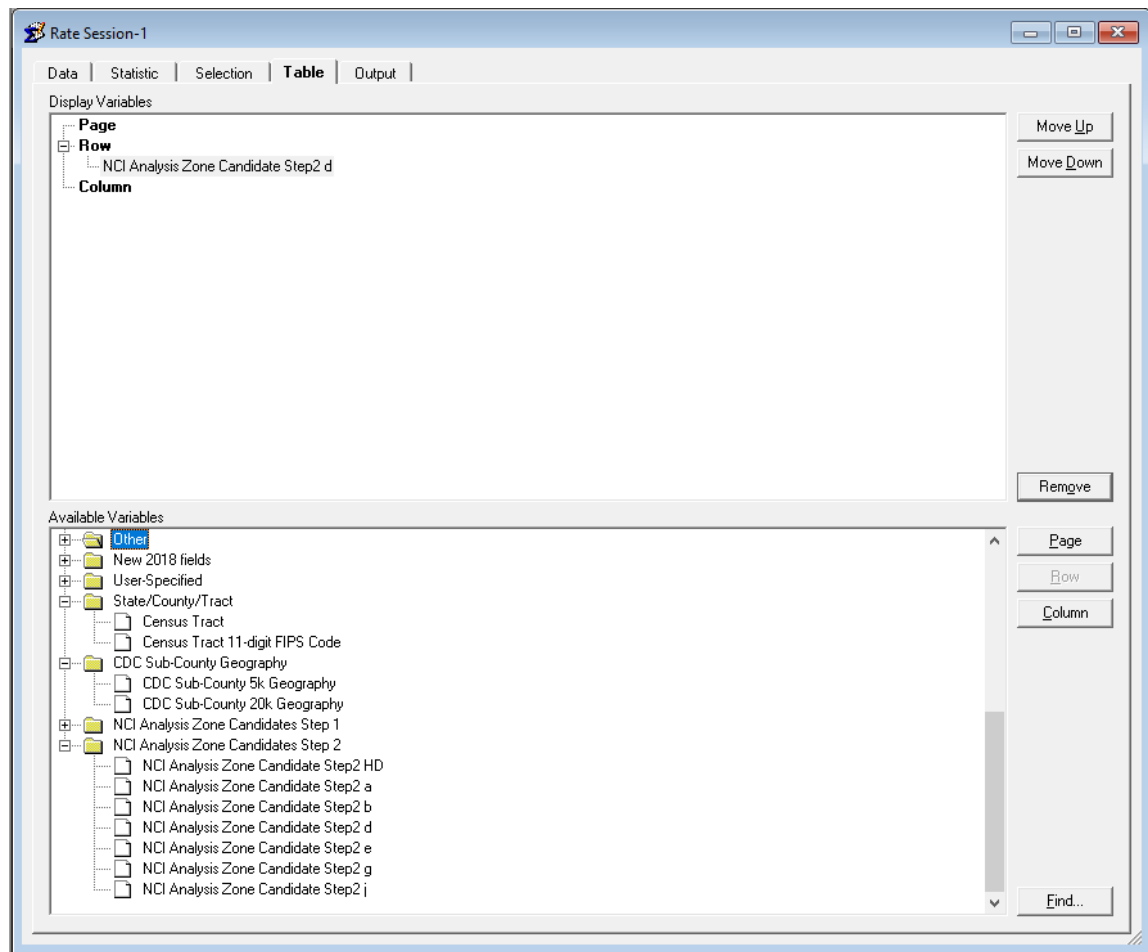
Other (Case Files)

{Site and Morphology, Behavior code ICD-O-3} = 'Malignant'
AND {Site and Morphology, Site recode ICD-O-3/WHO 2008} = ' Lung and Bronchus'

Edit...
Copy
Clear

☐ Select Only the First Matching Record for Each Person

Analysis by Census Tract in SEER*Stat



NCI/NAACCR Zones

	Rate	Count	Pop
A9001	58.0	360	614,297
A9002	64.4	215	532,300
A9003	36.4	187	498,009
A9004	52.7	283	580,356
A9005	64.4	520	654,934
A9006	35.8	159	487,747
A9007	46.7	322	557,124
A9008	39.6	192	539,249
A9009	53.7	336	544,036
A9010	41.5	224	522,819
A9011	49.5	327	618,285
A9012	60.7	303	601,200
A9013	59.7	540	678,271
A9014	62.8	364	571,677
B0001	29.1	160	482,184
B0005	61.9	614	617,768
B0008	59.9	438	483,370
B0012	41.8	308	540,010
B0013	34.7	178	479,866
B0016	38.5	280	643,690
B0018	25.5	99	557,417
B0021	38.2	293	758,423
B0023	52.7	398	477,590
B0024	64.8	418	542,727
B0025	59.8	316	478,098
B0026	49.7	417	732,542
Invalid Value(s)	~	1	0

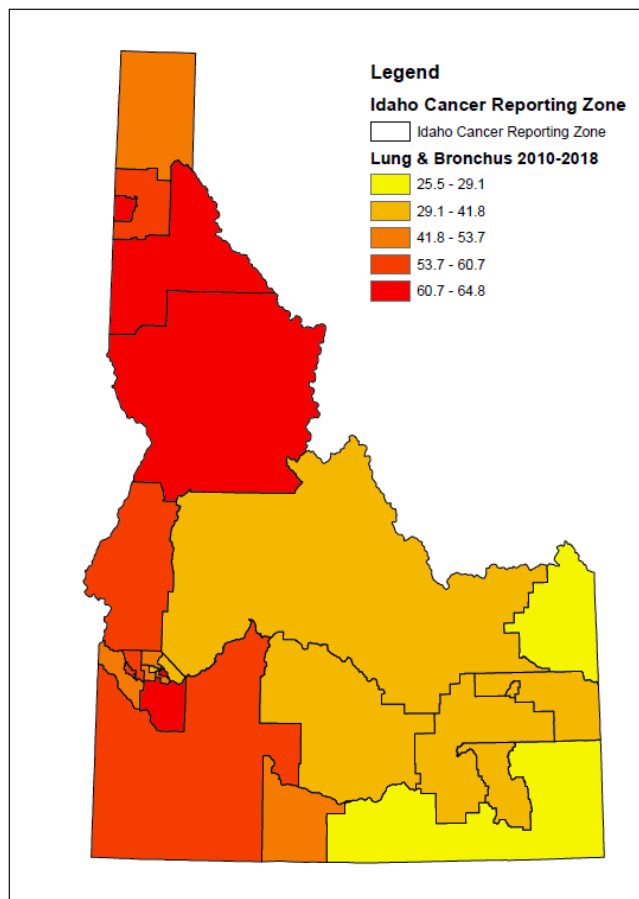
~ Rates are per 100,000 and age-adjusted to the 2000 US Std Population (19 age groups -
Statistic could not be calculated.

CDC 20k

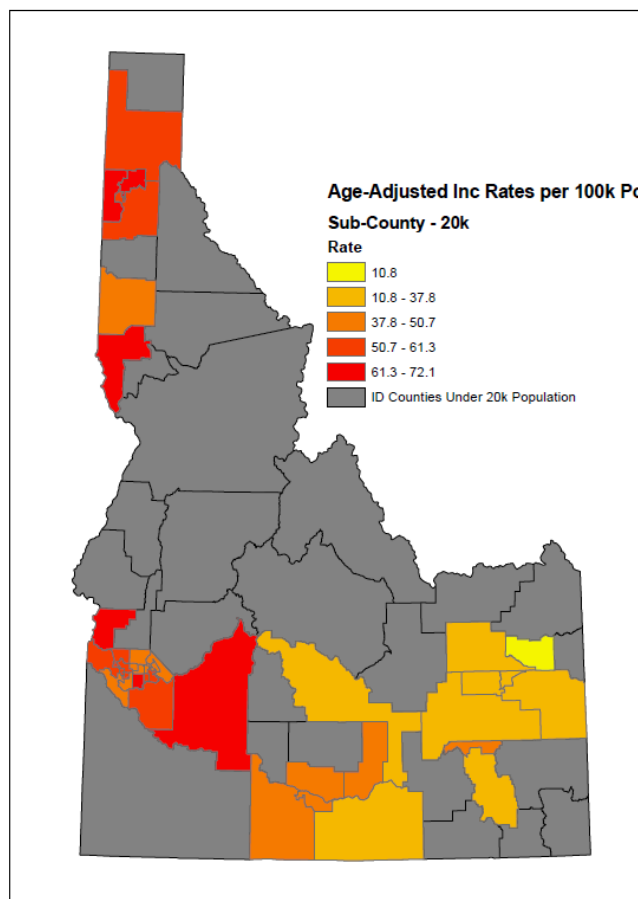
	Rate	Count	Pop
1600155	44.5	121	242,419
1600165	44.8	117	274,824
1600172	56.2	162	335,502
1600176	53.5	108	269,524
1600180	43.8	130	311,362
1600197	55.7	408	645,797
1600574	42.4	94	218,461
1600583	36.5	199	539,962
1601193	34.7	149	410,190
1601789	53.4	323	376,161
1601953	37.8	69	209,505
1601985	36.6	128	326,896
1602729	49.8	102	197,490
1602760	55.4	162	277,776
1602761	58.3	147	204,975
1602798	60.4	184	347,322
1603144	33.2	72	210,715
1603978	72.1	166	238,328
1605148	35.3	75	245,726
1605371	47.9	99	207,472
1605556	65.6	267	263,460
1605596	56.4	440	574,612
1605599	69.3	353	495,133
1605786	49.8	162	347,357
1606592	10.8	18	344,181
1606750	43.6	94	183,665
1606981	67.0	344	358,572
1607567	62.8	170	205,511
1608390	48.5	180	300,985
16001101	50.7	394	598,991
16001102	64.4	168	369,954
16001105	47.9	245	603,067
16013113	27.5	61	195,127
16019106	35.1	149	449,355
16027102	49.3	95	201,994
16027103	55.6	224	406,009
16027116	61.3	76	198,217
16083104	50.6	237	431,557
16001010335	45.1	83	196,119
Blank(s)	50.4	1,476	1,979,715
Invalid Value(s)	~	1	0

~ Rates are per 100,000 and age-adjusted to the 2000 US Std Population (19 age g
Statistic could not be calculated.

NCI/NAACCR Zone



CDC Sub-County 20k



Maps of
results

A large orange shape on the left side of the slide, consisting of a rectangle with a quarter-circle cutout on its right side.

Thank you!

- Questions & Answers to follow...
- 
- A yellow dashed line in the bottom right corner, composed of four short, curved segments that curve upwards and to the right.