

Readme for Cleaned_CDC_DATA.csv

Data Biography

This data was collected by the Center of Disease Control.

While reputable, the data itself has a bias toward the privacy of the deceased. Results that are too fine-grained risk leaking the private information of the deceased. If there is only one Asian family in a rural county with children, a search for deaths by race, age and county could reveal the family, and the Cause of Death in turn could reveal private information about them.

The data collected serves both the core mission of the CDC, and for sharing with other health organizations.

Unlike much of the data being collected, this is a population rather than a sample, and the particular data requested covers all counties in Alaska, Idaho, Oregon and Washington.

This data was collected using death certificates over the past 20 years.

The Data Comes with the Following Restrictions from the CDC

Data Use Restrictions:

The Public Health Service Act (42 U.S.C. 242m(d)) provides that the data collected by the National Center for Health Statistics (NCHS) may be used only for the purpose for which they were obtained; any effort to determine the identity of any reported cases, or to use the information for any purpose other than for health statistical reporting and analysis, is against the law. Therefore users will:

- Use these data for health statistical reporting and analysis only.
- For sub-national geography, do not present or publish death counts of 9 or fewer or death rates based on counts of nine or fewer (in figures, graphs, maps, tables, etc.).
- Make no attempt to learn the identity of any person or establishment included in these data.
- Make no disclosure or other use of the identity of any person or establishment discovered inadvertently and advise the NCHS Confidentiality Officer of any such discovery.

Confidentiality Officer
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20782
Telephone 888-642-4159
Email: nchsconfidentiality@cdc.gov

Sanctions for Violating Rules:

Researchers who violate the terms of the data use restrictions will lose access to WONDER and their sponsors and institutions will be notified. Researchers who are suspected of

violating the rules may be prevented from using WONDER until an investigation can be completed. Deliberately making a false statement in any matter within the jurisdiction of any department or agency of the Federal government violates 18 USC 1001 and is punishable by a fine of up to \$10,000 or up to 5 years in prison, or both.

Data Dictionary

Variable	Description	Data Type	Bias	Considerations
STATE	Abbreviation of State Name	String	None	Includes AK, ID, OR and WA
FIPS	FIPS code for the county	Integer	None	NA
YEAR	Year	Integer	None	NA
GENDER	Gender of deceased	String	Only M and F	NA
RACE	Race of deceased	String	When data is sparse, data not available on race	For many Idaho, Alaska, and some counties outside these states, must rely on “all races” category
AGE	Age by 10-year groups	String	None	NA
ICD_Chapter	Cause of death by Chapter Codes	String	Does not include comorbidities	Follow Link for Detailed Descriptions
DEATHS	Sum of number of deaths	Integer	None	Data for some races sometimes suppressed for privacy reasons
POPULATION	Sub-population in which deaths occurred	Integer	None	NA
ICD_CHAPTER_DESC	Longer description of ICD Chapter	String	None	Best to drop when deploying
COUNTY	Longform county name	String	None	Best to drop when deploying
Notes	Notes on data from CDC	String	None	Many notes concern crude deaths, which were dropped

Data Cleaning

To check original data and code used to clean, see:

<https://github.com/cmpear/tobacco>