About the United States Data on Causes of Death

By: Magali Barbieri[1]

Last updated: 15/04/2023

General

Cause-of-death data are collected as part of national vital statistics by the National Center for Health Statistics (NCHS), which was established in 1960. The NCHS was created as the result of a merge between the National Office of Vital Statistics, previously responsible for collecting and distributing vital statistics, and the National Health Survey. Population data (whether census or estimates) have been the responsibility of the U.S. Census Bureau since 1902. Data on population originate from population censuses conducted every 10 years. Census counts serve as the basis for producing annual and monthly population estimates for intercensal and postcensal periods, which have been published in *Current Population Reports: Population Estimates and Projections, Series P-25*, starting with the year 1970. Postcensal population estimates are frequently revised and intercensal population estimates are substituted for postcensal estimates as new census data become available. For recent years, population estimates are published on the Census Bureau web site (U.S. Census Bureau, Population Estimates Program). Vital statistics data are collected at the state level and centralized at the national level by the NCHS in order to assure uniformity. Since 1959, data on deaths are released by NCHS in the form of electronic public Detailed Mortality Files and Multiple Causes of Death Files (National Center for Health Statistics, 1959–, 1968–). These data include individual death records coded from death certificates. Data for the years since 1959 are directly available on the Centers for Disease Control and Prevention web site (CDC, http://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm).

Death statistics are published in the United States by the underlying cause of death, which is defined as the disease that directly leads to death. The exact definition adopted by the U.S. vital statistic system follows international recommendations. It is "(a) the disease or injury which initiated the train of events leading directly to death, or (b) the circumstances of the accident or violence which produced the fatal injury" (World Health Organization, 1975). Causes of death are classified according to the International Classification of Disease (ICD).

Territorial coverage

Until 1959, data on population and deaths refer to the territory of the United States, excluding Alaska and Hawaii. Since 1959, geographical coverage includes the 50 states and the District of Columbia. No other territorial change in coverage has occurred since 1959.

---

[1] This document is partly adapted from the Human Mortality Database Background and Documentation file, initially prepared for the United States by Ludmila Andreeva.

Part I – Vital statistics and population censuses

1. Death count data

Coverage and completeness

In the early years of the 20th century, vital statistics for the United States were based on data from those states admitted to the Death Registration Area, the number of which increased over time. This process was completed in 1933 with the admission of Texas. Given the legal requirements of death registration, the data on deaths are considered to be complete and of acceptable quality since 1933, explaining why the HMD mortality series for the United States start with this year.

Mortality data for the United States are confined to events registered in the United States (as defined in the previous section). The data do not cover vital events to U.S. residents occurring outside of the United States. If the death of a non-resident occurred in the United States, it was included in the U.S. vital statistics. For non-residents, the usual practice was to attribute the place of death as the place of residence. Since 1970, it is possible to identify deaths of non-residents and, consequently, to exclude them from tabulations. Therefore, for the period starting in 1970, death counts cover residents only.

Specific details

Deaths included in the Human Mortality Database and Human Cause-of-Death-Database have been tabulated from individual death records disseminated by NCHS in the form of Detailed Mortality Files (MDF) and Multiple Cause of Death Files (MCD) for all years 1959 and after. For all years when information on both the age at death and the date of birth are available (1989 and after), deaths were directly tabulated into Lexis triangles (see 1989-2013 for details). The detailed age at death was also used in cases where the age and birth cohort information were either inconsistent or missing as described below. For the years before 1989, deaths by single-year-of-age were allocated to Lexis triangles using all the information available. More specifically, for deaths below age one, the detailed age at death (provided in hours, minutes, weeks or months) was used to exactly determine the corresponding upper or lower Lexis triangle.

2. Population count data

Coverage and completeness

Data on population refer to the resident population of the United States. Although the Census Bureau provides estimates of census undercount rates—which generally vary from census to census and between population and broad age groups—we have made no corrections to the published population estimates.

3. Birth count data

Coverage and completeness

Like mortality statistics, due to the legal requirements of birth registration, the data on births are considered to be virtually complete and of a good quality since 1959, with some limitations for prior years. For the years 1933–1959, the Human Mortality Database (HMD) includes births adjusted for under-registration. The adjusted births for the period from 1935 to 1950 were published by NCHS (see NCHS, 1961, for years 1935 and 1940-1959, and NCHS, 1954 for years 1936–1939). For the years 1933–1934, the adjusted births were estimated by taking the published rates of under-reporting for both sexes combined and applying them to the unadjusted births by sex (NCHS, 1966). Both series of births are included in the raw data.

Birth data for the United States are confined to events registered in the United States. Births to legal residents of the U.S. that occurred in other parts of the world are excluded from published vital statistics. Prior to 1970, births to non-residents (that occurred in the U.S.) were included in the statistics (i.e., the *de facto* population), whereas for 1970 and thereafter, births to non-residents are excluded (i.e., the data refer only to the *de jure* population).

The official vital statistics counts in 1959 exclude births that occurred in Hawaii. In order to be consistent with the geographic coverage for population and mortality statistics, the counts have been adjusted to include births in Hawaii, which were available separately.

Part II – Information on CoD collection

4. Death certificate

Mortality statistics are derived from information reported on death certificates, which are collected by the civil registration system of each state. Some states produce their own statistics following national guidelines and quality control standards, while others send their certificates to NCHS, which processes them centrally. The certificates for each state have a slightly different content and format for demographic information, although a "standard" death certificate is recommended as a model by the central administration. The medical certification section of the certificate, which is used for the CoD information, is essentially the same for each state (Israel, Rosenberg, and Curtin, 1986). Although states can offer their own variations, most comply with the U.S. standard death certificate issued by the CDC's National Center for Health Statistics (NCHS). The agency last revised the form in 2003[2], and the first states to use it were California, Idaho, New York and Montana. By 2005, all US states had adopted the new standard certificate, which complies with the World Health Organization's International Classification of Diseases.

According to the National Association of Medical Examiners, about 20 percent of all death certificates are signed by a coroner or a medical examiner, compared to around 40% in 1970 and 50% in the 1940s (Goldman, 2018). Since autopsy rates have plummeted (to around 8%), death investigators now perform the majority of the nation's autopsies, which remain a vital barometer for revealing causes that might otherwise have been missed. Goldman explains that the reasons for such a rapid change are triple: "the lack of reimbursement for autopsies, the fear of disclosing mistakes that would lead to malpractice suits, and the belief that advances in medical technology, including but not limited to computed tomographic scans and magnetic resonance imaging, have made autopsies obsolete" (Goldman, 2018, p.2686). Others who can sign a death certificate include a primary physician, an attending physician, a non-attending physician, a medical examiner, a nurse practitioner, a forensic pathologist or a coroner; but this varies according to state law. In Texas, for example, a justice of the peace can sign. Typically, deaths have to be recorded with local health departments within 72 hours of the death, and to the state within five to seven days.

Once information is recorded by a clerk at the State Vital Statistics office, it is purchased by the National Center for Health Statistics (NCHS) – the division of the CDC responsible for compiling mortality data. The information is then used by a host of government agencies, as well as the private sector, to direct funding and future prevention policies.

---

[2] https://www.cdc.gov/nchs/data/dvs/death11-03final-acc.pdf

The NCHS has been collecting mortality data since 1979, but Dr. Robert Anderson, head of the division, says the states are often slow to deliver. And sometimes the agency lacks the funds to pay for all the information a state can provide. Budget cuts in 1995 forced the CDC to stop collecting data on the number of autopsies being performed. Although the process began again in 2003, it left an eight-year gap in vital records.

Quality checks on the cause of death data are performed on an ongoing basis by nosologists who code the diagnostic conditions at the state level as well as nosologists at NCHS who periodically review and verify data on a sample of death certificates submitted by the states (Center for Health Statistics, 1998). Errors may arise when an autopsy reveals that the original diagnosis on a death certificate was incorrect. This often happens when an autopsy has been requested because families, the courts, insurance companies or other interested parties are in dispute over how someone died. But getting that information amended on the death certificate, which is required by law, and reported back to the mortality division at the CDC can take months or even sometimes years. In some cases, it is never recorded as a matter of record.

5. Cause-of-death coding

As described by Israel and colleagues: *"In the initial sequence of steps through which the medical information on the death certificate is converted to statistical data, a medical coder translates [all] the entries on the certificate into standard codes of the International Classification of Diseases (ICD). Often, only the code for the underlying cause is recorded. In a properly completed certificate, with only one entry per line in Part I, the certifier's intended underlying cause is selected as the underlying cause for statistical tabulation. However, sometimes more than one condition is listed per line or an apparently erroneous causal sequence is listed. This type of improper certification may result from the lack of training of the physicians in certification practices or may be due to variation in the interpretations of terms such as underlying, immediate, or contributing causes. In 1976, only 55 per cent of death certificates in the U.S. had the underlying cause entered alone on the lowest used line in Part I"* (Israel, Rosenberg, and Curtin, 1986).

Israel further specifies that, for improperly completed certificates, medical coders or nosologists complete the information by hand whenever possible from the chain of events described on the certificates.

In 1968, computer application to CoD data in the U.S. began when an automated system (called ACME for "Automated Classification of Medical Entities") was implemented for the selection of the underlying CoD" (Israel, Rosenberg, and Curtin, 1986). ACME, which is still used by NCHS, requires the coding of all conditions shown on the medical certification. This innovation, which coincided with the implementation of a new ICD revision (ICD-8), led to experimentation with multiple cause data. Since then, NCHS has annually compiled data from all death certificate files in the United States and made these data available through a series of public-use tape files called "Multiple Cause of Death Files" (MCD Files). Each file contains a data record for all deaths processed by NCHS. Each data record contains the underlying cause, multiple cause, and demographic data for a death.

"Beginning with data year 1990, another computer system was implemented [in addition to ACME]. This system, called 'Mortality Medical Indexing, Classification, and Retrieval' (MICAR), automates the coding of the causes of death. The MICAR system is a major and logical step forward in the evolution of processing mortality data. MICAR takes advantage of the increasing capabilities of electronic data processing to produce information that is more consistently handled than manually processed information. In addition, MICAR provides more detailed information on the conditions reported on death certificate than is available in the ICD classification. Beginning with data year 1993, yet another computer system was implemented for automating cause-of-death coding. This system, called SuperMICAR, is an enhancement of the MICAR system, which allows for total literal entry of the multiple cause-of-death text as reported by the certifier. The numerical information produced by

SuperMICAR is then coded into the ICD scheme by the MICAR computer system and the underlying cause of death is selected according to international rules by the ACME system" (Center for Health Statistics, 1998). Indeed, "while the conditions listed in Part I should form a causal sequence initiated by the underlying cause listed on the lowest line, errors in properly completing the form occur regularly and a reselection of the underlying cause of death is done nationally 30-40% of the time. The decision to reselect an underlying cause other than that listed on the lowest used line in Part I is governed by a set of rules developed by WHO as part of the periodic revision of the International Classification of Disease and is incorporated, along with a complex set of decision tables, into [ACME]" (Wall et al., 1995, 2). "Records that are too complex or contradictory for SuperMICAR to process are coded manually before being submitted to ACME." (Center for Health Statistics, 1998).

6. Classifications in use and collected data

6.1 Classification changes

The United States has adopted the first revision to the ICD scheme in 1900 and has since implemented its successive revisions up to the 10th one, which is currently in use. Table 1 presents the periods of implementation of each ICD revision in the country.

Table 1. Periods of implementation of each ICD revision in the United States

ICD Revision	Years Covered
1st	1900-1909
2nd	1910-1920
3rd	1921-1929
4th	1930-1938
5th	1939-1948
6th	1949-1957
7th	1958-1967
8th	1968-1978
9th	1979-1998
10th	1999-present

Implementation of the ICD was a major advance in the comparison of mortality trends by cause across countries. However, the ICD is periodically revised to account for the identification of new diseases (HIV/AIDS, SIDS, or Alzheimer's to mention just a few recent examples), improved diagnosis and scientific accuracy of disease classification, and progress in medical knowledge. Though undoubtedly necessary, these revisions introduce disruptions in the series of mortality rates by cause, which are at times very significant. The most accurate method to overcome the issue of classification change is to adjust the series using the results of bridge-coding studies, also called dual-coding studies.

The National Center for Health Statistics (NCHS) has adapted to its own needs the International Classification of Disease, as published by the World Health Organization. The adjustments consist of disaggregating a number of disease categories for greater detail and precision. The adapted classification (ICDA) corresponds exactly to the original ICD at the 3-digit level, but a number of 4th-digit sub-divisions have been created to accommodate for the new categories introduced by the U.S. In a number of cases, it has been necessary to renumber an existing 4th-digit category in order to ensure a logical sequence between the additional subcategories. However, in the vast majority of cases, the renumbered sub-categories of the ICDA have an exact counterpart in the ICD. The most significant adjustment was that implemented for the 8th ICD revision (NCHS, 1977). Occasionally, an entirely new category is created by the United States to account for specific circumstances. For instance, after the

events of 9/11, NCHS added two categories coded U011 ("Terrorism – homicide") and U030 ("Terrorism – suicide").

6.2 Collected data

All cause-of-death data used in this project were downloaded from the NCHS web site[3]. Since a United States series is available in the Human Mortality Database, we are applying the cause-of-death fractions calculated from the MODICCOD-DIMOCHA project to the HMD death rates series, and this will allow us to obtain the final cause-specific death counts.

Table 2 provides an overview of collected data, including information about classification, age group format and data type.

---

[3] Mortality Multiple Cause Files for the United States, Centers for Disease Control and Prevention, National Center for Health Statistics, 1968-2018; downloaded from ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/Comparability/icd9_icd10 by Magali Barbieri on November 25, 2014 for years 1959-1988, December 1st, 2014 for years 1989-2003; March 16, 2015 for years 2004-2010; June 11, 2015 for years 2011-2013; November 12, 2020 for years 2014-2018.

Table 2. Overview of the data collected for the United States

Period	Classification	Number of items	List	Age group format	Data type
1968-1978	ICD-8-A	Standard ICD-8 categories plus additional U.S. specific sub-categories	4-digit items	Under 1 month, 1-11 months, 1 year, 2 years, 3 years, 4 years, 5-9 years, 10-14..., 95-99 years, 100 years and over, unknown	Individual electronic public-use data
1979-1998	ICD-9	Standard ICD-9 categories	4-digit items	Under 1 month, 1-11 months, 1 year, 2 years, 3 years, 4 years, 5-9 years, 10-14..., 95-99 years, 100 years and over, unknown	Individual electronic public-use data
1999-2018	ICD-10	Standard ICD-10 categories*	4-digit items	Under 1 month, 1-11 months, 1 year, 2 years, 3 years, 4 years, 5-9 years, 10-14..., 95-99 years, 100 years and over, unknown	Individual electronic public-use data
1996	Bridge-coding study for the transition from ICD-9 to ICD-10	Standard ICD-9 and ICD-10 categories	4-digit items	Under 1 month, 1-11 months, 1 year, 2 years, 3 years, 4 years, 5-9 years, 10-14..., 95-99 years, 100 years and over, unknown	Individual electronic public-use data

* Includes two additional U.S.-specific categories, namely U01.0 to U01.9 (Terrorism – Homicide), U02 (Sequelae of terrorism – Homicide), and U03.0 and U03.9 (Terrorism – Suicide)[4].

6.3 Data sources

All death counts for period 1999-2013 were accessed online from the Mortality Multiple Causes of Death Public Use Data File provided by the Division of Vital Statistics, National Center for Health Statistics, Centers for Disease Control.

6.4 Specific treatment of the raw data

There was no pre-treatment of the data downloaded from the internet.

7. Specific transition documents

The United States has been the first country to systematically implement bridge-coding studies, initially to document the transition from the 4th to the 5th ICD revision by systematically recoding all deaths from 1940 (Department of Commerce, 1944) and subsequently to document every other ICD change. However, results from the U.S. bridge-coding studies are only partially available up to the latest one, for which the complete electronic file of individual death records for 1996 is available, with each underlying cause classified according to both the 9th and 10th ICD revision schemes.

This 1996 file has been used to reconstruct the series of death counts by cause that were consistent over the periods when the two most recent ICD revisions were implemented (i.e., 1979-1998 and 1999

---

[4] See https://www.cdc.gov/nchs/icd/terrorism_code_appendix1.htm for more information.

to the present)[5]. The file provides information for nearly all deaths which occurred in the United States in 1996. Out of the original 2,318,212 records included in the 1996 mortality data file, 21,290 records (0.9%) were excluded for lack of ICD-10 information (missing value on the cause of death). These omissions were due to problems with the ICD-10 processing of these records, which occurred either because the original MICAR input files that were used for ICD-10 coding were lost or because the causes of death in the initially submitted records were either pending investigation or were later determined to be incorrect.

Part III – Reconstruction information

8. Reconstruction of coherent time series

During this first phase of the project, the reconstruction was conducted for the transition from the 9th to the 10th Revision of the ICD, which corresponds to the period 1979 through 2018. The next phase of the project will involve reconstructing the death count series for the transition from the 8th to the 9th Revision of the ICD.

As mentioned above, the results of the ICD-9 to ICD-10 bridge coding study conducted by NCHS in 1996 are available with full details, i.e., as an electronic file with individual level information for each death independently classified in both ICD-9 and ICD-10. Information on the sex and age of the deceased is also included.

Transition between ICD-9 and ICD-10

As a first step, the bridge-coding file was used to compute the transition coefficients for each 3-digit ICD-9 code into the 3-digit ICD-10 codes for both sexes combined, although for six separate age groups (i.e., infants, 1-19 years, 20-39 years, 40-59 years, 60-79 years, 80 years and over) Thus, we were able to take full advantage of both the size of the available sample and the availability of age information. These age-specific coefficients were then applied to all deaths in the corresponding age groups in the years when the 9th Revision of the ICD was in use, namely years 1979-1998. The resulting death count series transformed into the ICD-10 classification was linked to the aggregate death count by 3-digit ICD-10 code for all years when the 10th Revision of the ICD was in use, namely 1999-2018. For further examination, graphs of death counts by 3-digit ICD-10 codes for all years 1979-2018 were produced for all ages combined, as well as for each age group separately.

In a second step, manual redistribution was conducted for all series considered to be inconsistent (i.e., with a significant disruption from 1998 to 1999). An analysis of the disruption was conducted by examining the transition coefficients corresponding to the ICD-10 code under review, and a manual adjustment was implemented to correct the series. To the extent possible, the adjustment was made specifically for each age group. In addition to these manual adjustments to the transition coefficients, some *a posteriori* changes were implemented to redistribute deaths from one category to another when disruptions independent of the ICD-9 to ICD-10 transition were identified.

The main difficulty was that it was impossible to work at the association level, as was recommended for countries with no bridge-coding data. This is because there was a total of 6 associations: 5 of them with only one ICD-10 code and the last one with all remaining codes. Consequently, any change in the transition coefficient(s) for one ICD-10 category had the potential to disrupt every single ICD-10 series in the data file. In the future, I would thus recommend to everyone working with complete bridge-

---

[5] Public Use Multiple Cause of Death File on Comparability between ICD-9 and ICD-10 – a double-coded file based on the 1996 data year multiple cause of death file, Centers for Disease Control and Prevention, National Center for Health Statistics; downloaded from http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm by Magali Barbieri on March 17, 2008.

coding to start with the INED method and use the bridge-coding study only to guide the construction of the associations instead of trying to adjust the series reconstructed directly from the bridge-coding data.

9. Redistribution of ill-defined causes of death

Ill-defined causes of death were redistributed proportionally among other causes of death and accidents for the whole period.

References

Center for Health Statistics (1998). *Multiple Cause of Death in California, 1995*, Data Matters, Report Register no. DM98-0600, Department of Health Services, Health and Welfare Agency, State of California, 36 p.

Goldman, L. (2018). Autopsy 2018: still necessary, even if occasionally not sufficient. *Circulation*, *137*(25), 2686-2688.

Israel, R. A., Rosenberg, H. M., and Curtin, L. R. (1986). Analytical potential for multiple cause-of-death data. *American Journal of Epidemiology, 124*(2), 161-179.

National Center for Health Statistics (US). (2003). *Physicians' handbook on medical certification of death*. *The 2003 Revision*, US Department of Health and Human Services, Public Health Service, National Center for Health Statistics.

Wall, M., Huang, J., Oswald, J., andMcCullen, D. (2005). Factors associated with reporting multiple causes of death. *BMC Medical Research Methodology, 5*(4), 1-13.

List of acronyms

NCHS – National Center for Health Statistics

CDC – Centers for Disease Control

MCD – Multiple Cause-of-Death Files

Acknowledgments