Introduction

The linked birth/infant death data set (linked file) is now being released in two different formats - period data and birth cohort data. This documentation is for the 1998 period linked file. Beginning with 1995 data, the period linked files have formed the basis for all official NCHS linked file statistics (except for special cohort studies). Differences between period and birth cohort data are outlined below.

Period data - The numerator for the 1998 period linked file consists of all infant deaths occurring in 1998 linked to their corresponding birth certificates, whether the birth occurred in 1998 or 1997. The denominator file for this data set is the 1998 natality file, that is, all births occurring in 1998.

Birth cohort data - The numerator for the 1998 birth cohort linked file consists of deaths to infants born in 1998 whether the death occurred in 1998 or 1999. The denominator file is the 1998 natality file, that is, all births occurring in 1998. This file will be available about one year after the release of the period linked file.

The release of linked file data in two different formats allows NCHS to meet customer demands for more timely linked file data while still meeting the needs of data users who prefer the birth cohort format. While the birth cohort format has methodological advantages, it creates substantial delays in data availability, since it is necessary to wait until the close of the following data year to include all infant deaths to the birth cohort.

The 1998 period linked birth/infant death data set includes several data files. The first file includes all US infant deaths which occurred in the 1998 data year linked to their corresponding birth certificates, whether the birth occurred in 1998 or in 1997 - referred to as the numerator file. The second file contains information from the death certificate for all US infant death records which could not be linked to their corresponding birth certificates - referred to as the unlinked death file. The third file is the 1998 NCHS natality file for the US in compressed format, which is used to provide denominators for rate computations. These same three data files are also available for Puerto Rico, the Virgin Islands, and Guam.

Changes Beginning with the 1995 Data Year

In part to correct for known biases in the data, changes were made to the linked file beginning with the 1995 data year, and these changes are also effective for 1998 data. A weight has been added to the linked numerator file to correct in part

for biases in percent of records linked by major characteristics (see section on Percent of records linked below). The number of infant deaths in the linked file are weighted to equal the sum of the linked plus unlinked infant deaths by age at death and state. The formula for computing the weights is as follows:
number of linked infant deaths + number of unlinked infant deaths
number of linked infant deaths.

A separate weight is computed for each State of residence of birth and each age at death category (<1 day, 1-27 days, 28 days-1year). Thus, weights are 1.0 for states which link all of their infant deaths. The denominator file is not weighted. Weights have not been computed for the Puerto Rico, Virgin Islands, and Guam file.

An imputation for not-stated birthweight has been added to the data set, to reduce potential bias in the computation of birthweight-specific infant mortality rates. Basically, if birthweight is not-stated and the period of gestation is known, birthweight is assigned the value from the previous record with the same period of gestation, race, sex, and plurality. Imputed values are flagged. The addition of this imputation has reduced the percent of not-stated responses for birthweight from 3.63% to 1.29% in the numerator file, and from 0.12% to 0.05% in the denominator file, thus reducing (but not eliminating) the potential for underestimation when computing birthweight-specific infant mortality rates. The change from a birth cohort to a period format was discussed in detail on page one.

Comparisons of infant mortality data from the linked file with infant mortality data from the vital statistics mortality file

Although the time periods are the same, numbers of infant deaths and infant mortality rates by characteristics are not identical between the 1998 period linked file and the 1998 vital statistics mortality file.1 The differences can be traced to three different causes: 1) geographic differences; 2) additional quality control; and 3) weighting.

Geographic differences - To be included in the linked file for the 50 States and D.C., the birth and death must both occur inside the 50 States and D.C. In contrast, for the vital statistics mortality file, deaths which occur in the 50 States and D.C. to infants born inside and outside of the 50 States and D.C. are included. Similarly, to be included in the linked data file for Puerto Rico, the Virgin Islands, and Guam, the birth and death must both occur in Puerto Rico, the Virgin Islands or Guam. In contrast, for the vital statistics mortality file, deaths which occurred in Puerto Rico, the Virgin Islands, and Guam to infants born inside and outside of Puerto Rico, the Virgin Islands and Guam are included.

Additional quality control - The second reason for differences between the two files is that the linkage process subjects infant death records to an additional round of quality control review. Every year, a few records are voided from the file at this stage because they are found to be fetal deaths, deaths at ages greater than 1 year,

or duplicate death certificates.

Weighting - The third reason to the weighting procedures added to the 1995 and subsequent linked files. Beginning with 1995 data, linked file records are now weighted to compensate for the 2-3 percent of infant death records which could not be linked to their corresponding birth certificates. Although every effort has been made to design weights which will accurately reflect the distribution of deaths by characteristics, weighting may contribute to small differences in numbers and rates by specific variables between the linked file and the vital statistics mortality files.

In most cases, differences between numbers of infant deaths and infant mortality rates between the linked file and those computed from the vital statistics mortality file are negligible.

Methodology

The methodology used to create the national file of linked birth and infant death records takes advantage of two existing data sources:

1. State linked files for the identification of linked birth and infant death certificates; and

2. NCHS natality and mortality computerized statistical files, the source of computer records for the two linked certificates.

Virtually all States routinely link infant death certificates to their corresponding birth certificates for legal and statistical purposes. When the birth and death of an infant occur in different States, copies of the records are exchanged by the State of death and State of birth in order to effect a link. In addition, if a third State is identified as the State of residence at the time of birth or death, that State is also sent a copy of the appropriate certificate by the State where the birth or death occurred.

The NCHS natality and mortality files, produced annually, include statistical data from birth and death certificates that are provided to NCHS by States under the Vital Statistics Cooperative Program (VSCP). The data have been coded according to uniform coding specifications, have passed rigid quality control standards, have been edited and reviewed, and are the basis for official U.S. birth and death statistics.

To initiate processing, NCHS obtained matching birth certificate numbers from States for all infant deaths that occurred in their jurisdiction. We used this information to extract final, edited mortality and natality data from the NCHS natality and

mortality statistical files. Individual birth and death records were selected from their respective files and linked into a single statistical record, thereby establishing a national linked record file.

After the initial linkage, NCHS returned to the States where the death occurred computer lists of unlinked infant death certificates for follow up linking. If the birth occurred in a State different from the State of death, the State of birth identified on the death certificate was contacted to obtain the linking birth certificate. State additions and corrections were incorporated, and a final, national linked file was produced. Characteristics of the natality and mortality data from which the linked file is constructed are described in detail in the Technical Appendices and Addenda included in this document.

Characteristics of Unlinked File

For the 1998 linked file 467, or 1.6% of all infant death records could not be linked to their corresponding birth certificates. Unlinked records are included in a separate data file in this data set. The unlinked record file uses the same record layout as the numerator file of linked birth and infant death records. However, except as noted below, tape locations 1-210, reserved for information from the matching birth certificate, are blank since no matching birth certificate could be found for these records. The sex field (tape location 79) contains the sex of infant as reported on the death certificate, rather than the sex of infant from the birth certificate, which is not available. The race field (tape location 36-37) contains the race of the decedent as reported on the death certificate rather than the race of mother as reported on the birth certificate as is the case with the linked record file. The race of mother on the birth certificate is generally considered to be more accurate than the race information from the death certificate (see section on Comparison of race data from birth and death certificates in the Mortality Technical Appendix included in this documentation). Also, date of birth as reported on the death certificate is used to generate age at death. This information is used in place of date of birth from the birth certificate, which is not available.

Documentation table 6 shows counts of unlinked records by race and age at death for each State of residence. The user is cautioned in using table 6 that the race and residence items are based on information reported on the death certificate; whereas, tables 1-5 present data from the linked file in which the race and residence items are based on information reported on the birth certificate. (see section on Comparison of race data from birth and death certificates in the Mortality Technical Appendix included in this documentation).

Percent of Records Linked

The 1998 linked file includes 27,883 linked infant death records and 467 unlinked

infant death records. The linked file is weighted to the sum of linked plus unlinked records, thus the total number of weighted infant deaths by place of occurrence is 28,350. While the overall percent linked for infant deaths in the 1998 file is 98.4%, there are differences in percent linked by certain variables. These differences have important implications for how the data is analyzed.

Table 1 shows the percent of infant deaths linked by State of occurrence of death. While most States link a high percentage of infant deaths, linkage rates for some States are well below the national average. Note in particular the percent linked for California (95.9%), Maine (96.3%), New Mexico (96.1%), Ohio (94.6%) and Oklahoma (93.5%). When a high percentage of deaths remain unlinked, unweighted infant mortality rates computed for these States are underestimated. It is for this reason that weights were added to the file to correct for biases in the data due to poor data linkage for particular states.

The percent of infant deaths linked by race and age at death is shown in Table 2. In general, a slightly higher percentage of postneonatal (99.1%) than neonatal (98.2%) deaths were linked. The percent of records linked was 98.3% for infants of white and 98.5% for infants of black mothers. Variations in percent linked by underlying cause of death have also been noted (data not shown). While the weighting protocol has been designed to correct for possible bias due to variations in match rates by characteristics, no statistical method can correct perfectly for data limitations. Therefore, variations in the percent of records linked should be taken into consideration when comparing infant mortality rates by detailed characteristics.

Geographic classification

Geographic codes in this data set are based on the results of the 1990 census. Because of confidentiality concerns, only those counties and cities with a population size of 250,000 or more are separately identified in this data set. Users should refer to the geographic code outline in this document for the list of available areas and codes.

For events to be included in the linked file, both the birth and death must occur inside the 50 States and D.C. in the case of the 50 States and D.C. file; or in Puerto Rico, the Virgin Islands or Guam in the case of the Puerto Rico, Virgin Islands and Guam file. In tabulations of linked data and denominator data events occurring in each of the respective areas to nonresidents are included in tabulations that are by place of occurrence, and excluded from tabulations by place of residence. These exclusions are based on the usual place of residence of the mother. This item is contained in both the denominator file and the birth section of the numerator (linked) file. Nonresidents are identified by a code 4 in location 11 of these files.

Table 1. Percent of infant deaths linked by state of occurrence of death:
United States, 1998 linked file

United States	98.4%	Nebraska	100.0%
Alabama	100.0%	Nevada	97.4%
Alaska	98.1%	New Hampshire	100.0%
Arizona	97.5%	New Jersey	98.3%
Arkansas	97.7%	New Mexico	96.1%
California	95.9%	New York State	98.2%
Colorado	100.0%	New York City	99.2%
Connecticut	100.0%	North Carolina	99.7%
Delaware	100.0%	North Dakota	100.0%
District of Columbia	98.3%	Ohio	94.6%
Florida	99.7%	Oklahoma	93.5%
Georgia	100.0%	Oregon	100.0%
Hawaii	100.0%	Pennsylvania	98.0%
Idaho	100.0%	Rhode Island	100.0%
Illinois	99.0%	South Carolina	100.0%
Indiana	98.7%	South Dakota	100.0%
Iowa	100.0%	Tennessee	100.0%
Kansas	100.0%	Texas	98.0%
Kentucky	99.4%	Utah	98.2%
Louisiana	98.7%	Vermont	100.0%
Maine	96.3%	Virginia	99.4%
Maryland	99.2%	Washington	99.3%
Massachusetts	97.2%	West Virginia	99.4%
Michigan	98.6%	Wisconsin	100.0%
Minnesota	100.0%	Wyoming	100.0%
Mississippi	100.0%	Puerto Rico	99.6%
Missouri	97.5%	Virgin Islands	100.0%
Montana	98.6%	Guam	97.1%

Table 2. Percent of infant deaths linked by race and age at death: United States,
1998 linked file (Infant deaths are under 1 year; neonatal, under 28 days, and
postneonatal, 28 days-under 1 year)

	All races	White	Black
Infant	98.4%	98.3%	98.5%
Neonatal	98.2%	98.1%	98.3%
Postneonatal	99.1%	99.1%	99.0%

Demographic and Medical Classification

The documents listed below describe in detail the procedures employed for demographic classification on both the birth and death records and medical classification on death records. While not absolutely essential to the proper interpretation of the data for a number of general applications, these documents should nevertheless be studied carefully prior to any detailed analysis of demographic or medical (especially multiple cause) data variables. In particular, there are a number of exceptions to the ICD rules in multiple cause-of-death coding which, if not treated properly, may result in faulty analysis of the data.

A. Manual of the International Statistical Classification of Diseases, Injuries, and the Cause-of-Death, Ninth Revision (ICD-9) Volumes 1 and 2.

B. NCHS Instruction Manual Data Preparation Part 2a, Vital Statistics Instructions for Classifying the Underlying Cause-of-Death. Published annually.

C. NCHS Instruction Manual Data Preparation, Part 2b, Vital Statistics Instructions for Classifying Multiple Cause-of-Death. Published annually.

D. NCHS Instruction Manual Data Preparation, Part 2c, Vital Statistics ICD-9 ACME Decision Tables for Classifying Underlying Causes-of-Death. Published annually.

E. NCHS Instruction Manual Data Preparation, Part 2d, Vital Statistics NCHS Procedures for Mortality Medical Data System File Preparation and Maintenance, Effective 1985.

F. NCHS Instruction Manual Data Tabulation, Part 2f, Vital Statistics ICD-9 TRANSAX Disease Reference Tables for Classifying Multiple Causes-of-Death, 1982-85.

G. NCHS Instruction Manual Part 2g, Vital Statistics, Data Entry Instructions for the Mortality Medical Indexing, Classification, and Retrieval system (MICAR). Published annually.

H. NCHS Instruction Manual Part 2h, Vital Statistics, Dictionary of Valid Terms for the Mortality Medical Indexing, Classification, and Retrieval System (MICAR). Published annually.

I. NCHS Instruction Manual Data Preparation, Part 3a, Vital Statistics Classification and Coding Instructions for Live Birth Records. Published annually.

J. NCHS Instruction Manual Data Preparation, Part 4, Vital Statistics Demographic Classification and Coding Instructions for Death Records. Published annually.

K. NCHS Instruction Manual, Part 11, Vital Statistics Computer Edits for Mortality Data, Effective 1990.

L. NCHS Instruction Manual, Part 12, Vital Statistics Computer Edits for Natality Data, Effective 1993.

Copies of NCHS Instruction Manuals may be requested from the Chief, Data Preparation Branch, Division of Data Processing, National Center for Health Statistics, P.O. Box 12214, Research Triangle Park, North Carolina 27709.

In addition, the user should refer to the Technical Appendices of the Vital Statistics of the United States for information on the source of data, coding procedures, quality of the data, etc. The Technical Appendices for natality and mortality are part of this documentation package.

Cause-of-Death Data

Mortality data are traditionally analyzed and published in terms of underlying cause-of-death. The underlying cause-of-death data are coded and classified as described in the Mortality Technical Appendices. NCHS has augmented underlying cause-of-death data with data on multiple causes reported on the death certificate. The linked file includes both underlying and multiple cause-of-death data.

The multiple cause of death codes were developed with two objectives in mind. First, to facilitate etiological studies of the relationships among conditions, it was necessary to reflect

accurately in coded form each condition and its location on the death certificate in the exact manner given by the certifier. Secondly, coding needed to be carried out in a manner by which the underlying cause of death could be assigned through computer applications. The approach was to suspend the linkage provisions of the ICD for the purpose of condition coding and code each entity with minimum regard to other conditions present on the certification. This general approach is hereafter called entity coding.

Unfortunately, the set of multiple cause codes produced by entity coding is not conducive to a third objective -- the generation of person-based multiple cause statistics. Person-based analysis requires that each condition be coded within the context of every other condition on the same certificate and modified or linked to such conditions as provided by ICD-9. By definition, the entity data cannot meet this requirement since the linkage provisions distort the character and placement of the information originally recorded by the certifying physician.

Since the two objectives are incompatible, NCHS has chosen to create from the original set of entity codes a new code set called record axis multiple cause data. Essentially, the axis of classification has been converted from an entity basis to a record (or person) basis. The record axis codes are assigned in terms of the set of codes that best describe the overall medical certification portion of the death certificate.

This translation is accomplished by a computer system called TRANSAX (translation of axis) through selective use of traditional linkage and modification rules for mortality coding. Underlying cause linkages which simply prefer one code over another for purposes of underlying cause selection are not included. Each entity code on the record is examined and modified or deleted as necessary to create a set of codes which are free of contradictions and are the most precise within the constraints of ICD-9 and medical information on the record. Repetitive codes are deleted. The process may (1) combine two entity axis categories together to a new category thereby eliminating a contradiction or standardizing the data; or (2) eliminate one category in favor of another to promote specificity of the data or resolve contradictions. The following examples from ICD-9 illustrate the effect of this translation:

Case 1: When reported on the same record as separate entities, cirrhosis of liver and alcoholism are coded to 5715 (cirrhosis of liver without mention of alcohol) and 303 (alcohol dependence syndrome). Tabulation of records with 5715 would on the surface falsely imply that such records had no mention of alcohol. A referable codification would be 5712 (alcoholic cirrhosis of liver) in lieu of both 5715 and 303.

Case 2: If "gastric ulcer" and "bleeding gastric ulcer" are reported on a record they are coded to 5319 (gastric ulcer, unspecified as acute or chronic, without mention of hemorrhage or perforation) and 5314 (gastric ulcer, chronic or unspecified, with hemorrhage). A more concise codification would be to code 5314 only since the 5314 shows both the gastric ulcer and the bleeding.

Entity Axis Codes

The original conditions coded for selection of the underlying cause of death are reformatted and edited prior to creating the public-use tape. The following paragraphs describe the format and application of entity axis data.

Format - Each entity-axis code is displayed as an overall seven byte code with subcomponents as follows:

1. Line indicator: The first byte represents the line of the certificate on which the code appears. Six lines (1-6) are allowable with the fourth and fifth denoting one or two written in "due to"s beyond the three lines provided in Part I

of the U.S. standard death certificate. Line "6" represents Part II of the certificate.

2. Position indicator: The next byte indicates the position of the code on the line, i.e., it is the first (1), second (2), third (3),... eighth (8) code on the line.

3. Cause category: The next four bytes represent the ICD-9 cause code.

4. Nature of injury flag: ICD-9 uses the same series of numbers (800-999) to indicate nature of injury (N codes) and external cause codes (E codes). This flag distinguishes between the two with a one (1) representing nature of injury codes and a zero (0) representing all other cause codes.

A maximum of 20 of these seven byte codes are captured on a record for multiple-cause purposes. This may consist of a maximum of 8 codes on any given line with up to 20 codes distributed across three or more lines depending on where the subject conditions are located on the certificate. Codes may be omitted from one or more lines, e.g., line 1 with one or more codes, line 2 with no codes, line 3 with one or more codes.

In writing out these codes, they are ordered as follows: line 1 first code, line 1 second code, etc. ----- line 2 first code, line 2 second code, etc. ----- line 3 ----- line 4 ----- line 5 ----- line 6. Any space remaining in the field is left blank. The specifics of locations are contained in the record layout given later in this document.

Edit - The original conditions are edited to remove invalid codes, reverify the coding of certain rare causes of death, and assure age/cause and sex/cause compatibility. Detailed information relating to the edit criteria and the sets of cause codes which are valid to underlying cause coding and multiple cause coding are provided in Part 11 of the NCHS Vital Statistics Instruction Manual Series.

Entity axis applications - The entity axis multiple cause data is appropriate to analyses which require that each condition be coded as a stand alone entity without linkage to other conditions and/or require information on the placement of such conditions in the certificate. Within this framework, the entity data are appropriate to the examination of etiological relationships among conditions, accuracy of certification reporting, and the validity of traditional assumptions in underlying cause selection.

Additionally, the entity data provide in certain categories a more detailed code assignment which is linked out in the creation of record axis data. Where such detail is needed for a study, the user should selectively employ entity data. Finally, the researcher may not wish to be bound by the assumptions used in the

axis translation process preferring rather to investigate hypotheses of his own predilection.

By definition, the main limitation of entity axis data is that an entity code does not necessarily reflect the best code for a condition when considered within the context of the medical certification as a whole. As a result certain entity codes can be misleading or even contradict other codes in the record. For example, category 5750 is titled "Acute cholecystitis without mention of calculus". Within the framework of entity codes this is interpreted to mean that the codable entity itself contained no mention of calculus rather than that calculus was not mentioned anywhere on the record. Tabulation of records with a "5750" as a count of persons having acute cholecystitis without mention of calculus would therefore be erroneous. This illustrates the fact that under entity coding the ICD-9 titles cannot be taken literally. The user must study the rules for entity coding as they relate to his/her research prior to utilization of entity data. The user is further cautioned that the inclusion notes in ICD-9 which relate to modifying and combining categories are seldom applicable to entity coding (except where provided in Part 2b of the Vital Statistics Instruction Manual Series).

In tabulating the entity axis data, one may count codes with the resultant tabulation of an individual code representing the number of times the disease(s) represented by the code appears in the file. In this kind of tabulation of morbid condition prevalence, the counts among categories may be added together to produce counts for groups of codes. Alternatively, subject to the limitations given above, one may count persons having mention of the disease represented by a code or codes. In this instance it is not correct to add counts for individual codes to create person counts for groups of codes. Since more than one code in the researcher's interest may appear together on the certificate, totaling must account for higher order interactions among codes. Up to 20 codes may be assigned on a record; therefore, a 20-way interaction is theoretically possible. All totaling must be based on mention of one or more of the categories under investigation.

Record Axis Codes

The following paragraphs describe the format and application of record-axis data. Part 2f of the Vital Statistics Instruction Manual Series describes the TRANSAX process for creating record axis data from entity axis data.

Format - Each record (or person) axis code is displayed in five bytes. Location information is not relevant. The Code consists of the following components:

1. Cause category: The first four bytes represent the ICD-9 cause code.

2. Nature of injury flag: The last byte contains a 0 or 1 with the 1 indicating that the cause is a nature of injury category.

Again, a maximum of 20 codes are captured on a record for multiple cause purposes. The codes are written in a 100-byte field in ascending code number (5 bytes) order with any unused bytes left blank.

Edit - The record axis codes are edited for rare causes and age/cause and sex/cause compatibility. Likewise, individual code validity is checked. The valid code set for record axis coding is the same as that for entity coding.

Record axis applications - The record axis multiple cause data set is the basis for NCHS core multiple cause tabulations. Location of codes is not relevant to this data set and conditions have been linked into the most meaningful categories for the certification. The most immediate consequence for the user is that the codes on the record already represent mention of a disease assignable to that particular ICD-9 category. This is in contrast to the entity code which is assigned each time such a disease is reported on two different lines of the certification. Secondly, the linkage implies that within the constraints of ICD-9 the most meaningful code has been assigned. The translation process creates for the user a data set which is edited for contradictions, duplicate codes, and imprecisions. In contrast to entity axis data, record axis data are classified in a manner comparable to underlying cause of death classification thereby facilitating joint analysis of these variables. Likewise, they are comparable to general morbidity coding where the linkage provisions of ICD-9 are usually utilized. A potential disadvantage of record axis data is that some detail is sacrificed in a number of the linkages.

The user can take the record axis codes as literally representing the information conveyed in ICD-9 category titles. While knowledge of the rules for combining and linking and coding conditions is useful, it is not a prerequisite to meaningful analysis of the data as long as one is willing to accept the assumptions of the axis translation process. The user is cautioned, however, that due to special rules in mortality coding, not all linkage notes in ICD-9 are utilized. (See Part 2f of the Vital Statistics Instruction Manual Series.)

The user should proceed with caution in using record axis data to count conditions as opposed to people with conditions since linkages have been invoked and duplicate codes have been eliminated. As with entity data, person based tabulations which combine individual cause categories must take into account the possible interaction of up to 20 codes on a single certificate.

In using the NCHS multiple cause data, the user is urged to review the information in this document and its references. The instructional material does change from year to year and revision to revision. The user is cautioned that coding of specific ICD-9 categories should be checked in the appropriate instruction manual. What may appear on the surface to be the correct code by ICD-9 may in fact not be correct as given in the instruction manuals.

If on the surface it is not obvious whether entity axis or record axis data should be employed in a given application, detailed examination of Part 2f of the Vital Statistics Instruction Manual Series and its attachments will probably provide the necessary information to make a decision. It allows the user to determine the extent of the trade-offs between the two sets of data in terms of specific categories and the assumptions of axis translation. In certain situations, a combination of entity and record axis data may be the more appropriate alternative.

1 see: Murphy SL. Deaths: Final Data for 1998. National vital statistics report; vol. 48 no. 11. Hyattsville, Maryland: National Center for Health Statistics. 2000.