the NIMS project was unable to quantify this potential source of bias.

Conclusions

The record linkage process both within and between States could be improved to attain more complete linkage of birth and infant death certificates. Although unlinked infant death certificates had little effect on the infant mortality risks overall at the State and national levels, the underreporting of births may be different for various subgroups, such as low birth weight infants. When linked record data are used, those persons doing epidemiologic studies and developing programs and policies should consider carefully the quality of record linkage.

References

1. Office of the Assistant Secretary for Health and Surgeon General: Healthy people: the Surgeon General's report on health promotion and disease prevention. DHEW (PHS) Publication No. 79-55071. U.S. Government Printing Office, Washington, DC, 1979.
2. Kleinman, J. C., et al.: A comparison of 1960 and 1973-1974 early neonatal mortality in selected states. Am J Epidemiol 108: 454-469 (1978).
3. Armstrong, R. J.: A study of infant mortality from linked records by birth weight, period of gestation, and other variables. Vital Health Stat [20], No. 12. National Center for Health Statistics, Rockville, MD, 1972.
4. Philip, A. G. S., et al.: Neonatal mortality risk for the eighties: the importance of birth weight/gestational age groups. Pediatrics 68: 122-130 (1981).
5. Susser, M., Marolla, F. A., and Fleiss, J.: Birth weight, fetal age, and perinatal mortality. Am J Epidemiol 96: 197-204 (1972).
6. Hogue C. J. R., et al.: Overview of the National Infant Mortality Surveillance (NIMS) project—design, methods, results. Public Health Rep 102: 126-138, March-April 1987.
7. Chase, H. C.: A study of infant mortality from linked records: methods of study and registration aspects: United States: 1960 live-birth cohort. Vital Health Stat [20], No. 7. National Center for Health Statistics, Rockville, MD, 1970.
8. Frost, F., and Shy, K. K.: Racial differences between linked birth and infant death records in Washington State. Am J Public Health 70: 974-976 (1980).
9. Williams, R. L., and Chen, P. M.: Special article: identifying the sources of the recent decline in perinatal mortality rates in California. N Engl J Med 306: 207-214, Jan. 28, 1982.
10. Strauss, L. T., et al.: Experiences with linked birth and infant death certificates from the NIMS project. Public Health Rep 102: 204-210, March-April 1987.
11. McCarthy, B. J., et al.: The underregistration of neonatal deaths: Georgia 1974-77. Am J Public Health 70: 977-982 (1980).

Experiences with Linked Birth and Infant Death Certificates from the NIMS Project

LILO T. STRAUSS, MA
MARY ANNE FREEDMAN, MS
NITA GUNTER, MA
EVE POWELL-GRINER, PhD
JACK C. SMITH, MS

Two of the authors are with the Center for Health Promotion and Education, Centers for Disease Control (CDC), Atlanta, GA. Mr. Smith is Chief and Ms. Strauss is Mathematical Statistician, Research and Statistics Branch, Division of Reproductive Health.

Ms. Freedman is Director, Division of Public Health Statistics, Vermont Department of Health, Burlington. Ms. Gunter is with the Bureau of Information Resources, Public Health Statistics, Jackson, MS. Dr. Powell-Griner is with the Mortality Statistics Branch, Division of Vital Statistics, National Center for Health Statistics, Hyattsville, MD.

Other contributors from CDC's Research and Statistics Branch are Jeanne C. Gilliland, J. Patrick Whitaker, and Evelyn L. Finch, who worked on systems design and assisted with computer programming in aggregating data from 53 vital statistics reporting areas; Sara W. Gill and Merrell Ramick, who assisted in preparing the data for processing; and Phyllis A. Wingo, who coordinated the data management.

This research was supported in part by the National Institute for Child Health and Human Development, the Health Resources and Services Administration, and the National Center for Health Statistics, all agencies of the Public Health Service.

Tearsheet requests to NIMS Coordinator, DRH, CHPE, Centers for Disease Control, Atlanta, GA 30333.

Synopsis

The National Infant Mortality Surveillance (NIMS) project aggregated data provided by 53 vital statistics reporting areas—50 States, New York City, the District of Columbia, and Puerto Rico (subsequently called States)—from their files of linked birth and death certificates and compared individual States' total infant mortality experiences for the 1980 birth cohort by age at death, race, birth weight, and plurality. Therefore, it was essential to achieve maximum uniformity among the separate data sets and to specify when this uniformity could not be obtained.

In working with these multiple sources, we identified five key issues that relate to data from linked birth and death certificates: (a) *Variations in definitions of variables are often embedded in data that have been gathered from several independent sources. (For NIMS, the sources were 53 reporting areas and the National Center for Health Statistics.)* (b) *Variations in States' linking procedures—these are based on an individual State's primary purpose for linking the data—affect the completeness and comparability of the 1980 resident birth cohorts used for NIMS.* (c) *Variations in the recording of some pregnancy outcomes as fetal deaths or live births are known to be a problem in vital statistics data that particularly affects data for events among infants weighing less than 500 g at birth.* (d) *Ambiguities occur frequently in unknowns or zero values. For NIMS this effect was most pronounced for the pregnancy history variables. Examination of the values reported for unknown or zero categories helps in uncovering problems with and improving quality of data.* (e) *Analysis from a new perspective may reveal unexpected data problems. These problems tend to surface only during a reexamination of underlying data that is prompted by unusual findings.*

Continued alertness to these issues may improve further the quality of data in files of linked birth and death certificates and assure the integrity of analysis based on these data.

THE NATIONAL INFANT MORTALITY Surveillance (NIMS) project was a large undertaking that required the coordinated efforts of the Centers for Disease Control (CDC) and States' health statistics offices. In any project of this magnitude, problems and issues inevitably emerge. In this paper we present the most common of these issues with the hope that a better understanding of the data base will result and that potential pitfalls will be avoided when working with linked birth and death data in the study of infant mortality.

From its inception, the NIMS project faced the potentially conflicting goals of collecting uniform data from the States and of simplifying the data request. We knew that States might have used different definitions and categories for collecting, coding, and tabulating data on birth and death certificates and in establishing their linked data files. We anticipated some of the problems for obtaining comparable data, and other problems emerged as we communicated with States while they were preparing the data requested by CDC. Still other problems emerged as the data were reviewed at CDC after initial submission.

Methods

The methods of the NIMS project, including data collection and evaluation, are described elsewhere (*1–3*). In brief, 53 vital statistics reporting areas (subsequently referred to as States) participated in the project: 50 States, New York City, the District of Columbia, and Puerto Rico. All 53 reporting areas linked birth and death certificates for infants who were born alive in 1980 and who died within the first year of life in 1980 or 1981. States provided CDC with the number of infant deaths according to birth weight, age at death, and other infant and maternal characteristics. CDC generated corresponding numbers of births from the computer tape of 1980 natality records produced by the National Center for Health Statistics (NCHS), with exceptions for Maine and New Mexico as previously described (*1*).

At the time of the data request, CDC provided supportive materials to the States to improve uniformity of data collected from different States and to evaluate and document the quality of the linked data. CDC provided each State these materials with its initial request for data: general instructions (including an outline of possible problems in completing data tables and notes for programmers), specific definitions and valid ranges for maternal and infant characteristics used in the data tables, examples of tabular formats, a definitions checklist that asked for each variable whether the State was able to use the recommended definition (and directions to check with CDC before using a modification of any definition), and a sample SAS (*4*) program using CDC-specified definitions and groupings.

In addition to examining the definitions checklist, we reviewed 1980 birth and death certificate forms for each State to ensure that the data submitted could comply with CDC's definitions. We also checked values in zero and unknown categories, edited for internal consistency of data, and reviewed the computer program (when available) that States used in preparing NIMS data.

We also dealt with approximately 500 specific

telephone queries—some initiated by CDC and some by State personnel—for clarification and modification. This process made us sensitive to unforeseen issues and the way in which various States handled these issues. Frequently, only through indepth communication between CDC and a State did each realize that there was an issue inherent in the data that needed to be discussed. The same cross-fertilization of methods and ideas occurred among participants at the several NIMS Conference workshops. Many issues were addressed, and several participants elaborated on how they resolved a problem area, thus enlightening other workshop attendees.

Issues

All of the methods mentioned helped us to identify five key issues that relate to data from linked birth and infant death certificates:

- Variations in variable definitions are often embedded in data gathered from several independent sources. For NIMS the sources were 53 reporting areas and the NCHS.
- Variations in States' linking procedures, based on varying primary purposes for individual States, affect the completeness and comparability of the 1980 resident birth cohorts used for NIMS.
- Variations in the recording of pregnancy outcome as fetal death or live birth are known to be a problem in vital statistics data that particularly affects data for events of less than 500 g birth weight.
- Unknown or zero values frequently hide ambiguities, but examining those values is helpful in uncovering data problems and improving data quality.
- Analysis may reveal unexpected data problems. These problems tend to surface only through another examination of underlying data prompted by unusual findings.

We discuss each of these issues subsequently.

Variations in variable definitions. We used U.S. Standard Certificate of Live Birth and NCHS definitions and categorizations for all variables (except for the type of delivery, when listed as a separate item on a State's birth certificate). In spite of using the standard NCHS definitions, 22 States initially could not provide all the requested variables as defined, ranging from 13 States unable to comply for one variable to 1 State unable to comply for 6 variables. The least uniformly defined variables were number of other terminations of pregnancy (20 or more weeks' gestation) and number of previous live births. Other variables for which uniformity could not be achieved were race of infant (three States), gestational age at birth (three States), number of prenatal care visits (three States), maternal education (three States), month prenatal care began (three States), sex (two States), and birth weight (one State). These include States that did not include the variable either on their certificate or in their State data base. Interestingly, most States were not aware they were deviating from NCHS definitions. Approximately one-third of the States indicated, for at least one variable, that they were able to use the recommended definition as we requested, when, in fact, detailed work with the data at CDC revealed discrepancies between the recommended and the utilized definitions.

Definitions and categorizations of race of infant and birth weight had to comply with NIMS specifications as a minimum requirement for project participation. Therefore, we expected no surprises concerning these variables. We used the NCHS guidelines to define race of infant and collected data for whites, blacks, and all races, with infants of other or unknown races not tabulated separately. Although each State indicated in the definitions checklist that they complied with the NCHS definition, we found three deviations in the definition of race of infant (*1*); however, we were able to use all data. The definition of birth weight categories was critical for NIMS. All States except one were able to use our groupings (*1*), and NIMS accommodated data from that State with minimal limitation.

Several States indicated on the definitions checklist that they could use CDC's definitions of pregnancy history variables without deviation, but further examination indicated they could not. The deviation in definition of previous live births consisted of the States' inability to separate the zero and unknown categories, discussed subsequently under "ambiguities with unknowns or zero values." The second pregnancy history variable, other terminations of 20 or more weeks' gestation, presented problems because several components are included in the variable and because the source for this data item does not appear uniformly on States' birth certificates. We specified that stillbirths and both spontaneous and induced abortions on and after 20 weeks' gestation be included in the data. NIMS data include variations from

our recommended definitions when these variations allowed us to use best available data without significant distortion. For example, we included the data in the NIMS categories when one State reported fetal deaths at 20 or more weeks but did not include induced abortions, two other States' birth certificates listed only "miscarriages" or "were born dead (20 weeks or more pregnancy)," and another State could provide "other terminations after 16 weeks." When several other States' birth certificates listed only "fetal deaths at any time after conception" or "other terminations" with gestation not specified, we did not include these data in the detailed categories but used them in the aggregate under "unknown."

Variations in States' linking procedures. Published NIMS data are based on aggregates of data from each State's 1980 birth cohort (*2*) and therefore are constrained by the completeness and comparability of these cohorts. We found that the linking procedures used in generating birth and death cohorts varied widely. All States begin the linkage process by attempting to find a matching birth record for each infant death. About half of the States end with birth cohorts and half with death cohorts for each year, with several States maintaining both. The most complete State files contain both deaths and births for both occurrence and residence of each of these events in the State, with the files being updated as records are received. The most restricted files link a resident death file (excluding deaths which occur out of State) with a resident birth file, or an occurrence death file with an occurrence birth file, and have an early cutoff date for entering records into the linked file. Most linking procedures fall between these two extremes.

States varied in use of birth or death cohort, inclusion of nonresidents at death, inclusion of out of State reports in the statistical file, and cutoff dates. Also, some linked files include only linked records, and most files maintained by birth cohorts include all births for a given year with infant death records linked to applicable birth records. Some States may reflect an artificially lower infant mortality experience compared with States whose linked files were more complete: States with lower mortality might have excluded from their linked resident birth cohort file (*a*) deaths of infants who were not residents at death, (*b*) births or deaths reported after the cutoff date for their file, or (*c*) deaths reported through the Interstate Vital Records Exchange System (*5*). Although we estimated each State's completeness of reporting by asking for detailed data on numbers of unlinked death certificates, these estimates are of limited use because the linking process itself determines whether a death certificate is included in the death file to be linked. A detailed evaluation of reported unlinked infant death certificates appears elsewhere in this issue (*3*). Independent evaluation allowed us to estimate that NIMS data include approximately 95 percent of infant deaths (*1–3*). In addition to deaths not included because of the variation in the States' linking procedures, some deaths are not captured when aggregating individual State data (NIMS) rather than having a nationally linked data file (proposed NCHS process (*6*)). For example, the death certificate of an infant who is a resident of State A and is born at a tertiary center across State lines in State B, then *moves* to and later dies in State C, would not show up in the resident birth cohort of either States A, B, or C even with optimal interstate exchange of certificates. This situation particularly affects the counts of States with tertiary care centers near a State border (*3*).

Variations in recording pregnancy outcome as fetal death or live birth. Another issue for NIMS (and all other birth-weight-specific analyses) surfaced in its full importance through participant exchange at the NIMS Conference, namely, the unreliability of reported live birth and infant death data for infants of birth weights less than 500 g and its consequences on analysis of mortality of very low birth weight infants.

A major factor underlying this unreliability is the classification of pregnancy outcomes of less than 500 g. Is it a live infant who dies (infant death) or is it a fetal death? Although uniform definitions presumably are used, the attending physician decides which certificate to file, and the physician may take the easier route of filing one certificate (fetal death) rather than two (live birth and infant death) for infants who live only briefly. Recognizing this factor, some States reported that they search fetal death records when they cannot find a matching birth certificate for an infant death.

Another issue pertaining to low birth weights arose when we discovered that some birth weights of less than 1 pound (less than 500 g) recorded for one State had been converted (when changed from pounds and ounces to grams) to the unknown birth weight category.

The table illustrates the impact of these variations in reporting data about birth weights less

Live births and deaths of infants with reported birth weights less than 500 g expressed as percent of all live births and of deaths, U.S. vital statistics reporting areas[1]

	Births			Deaths		
	Range	Median percent	U.S. total (= mean)	Range	Median percent	U.S. total (= mean)
All infants:						
All races	0.02–0.36	0.09	0.10	1.51–14.72	7.56	8.09
Whites	0 –0.12	0.07	0.07	1.51–11.94	6.58	6.93
Blacks	0 –0.97	0.20	0.23	0 –35.48	11.05	11.62
Single-delivery infants:						
All races	0.01–0.35	0.07	0.08	0.70–14.55	6.38	7.09
Multiple-delivery infants:						
All races	0 –2.44	0.97	1.07	0 –51.47	15.40	17.24

[1] Areas with known birth weight reporting problems excluded from the ranges.

than 500 g. The table lists the range, median, and means of percents of events of birth weights less than 500 g excluding extremes of States with known problems. Although small numbers may affect the stability of some high and low values, clearly more than biologically plausible variations are involved in these ranges.

Because of the uncertainty of the data concerning birth weights less than 500 g, we decided to present aggregate data in the NIMS Report both including and excluding events of birth weights less than 500 g. In making this decision, we considered the concerns expressed in the experiences that several States' representatives shared. At the same time, we responded to the need for full information voiced by the representatives. For instance, neonatal mortality risks are presented for both the less than 1,500 g and the 500 to 1,499 g categories.

Ambiguities with unknowns or zero values. In NIMS, as in all data collection, we encountered frequent ambiguities with data reported as either unknown or zero, in spite of conscientious efforts to be clear in data collection, coding, and programming. Errors with respect to unknown and zero values can occur when information is initially entered on certificates; when data are abstracted from the certificates; when coding for the births, deaths, or linked files; and when computer programs are written to create files at the State level or to create files for special projects such as NIMS. We encountered *all* of these situations, and we are convinced that the relative unreliability of unknown and zero values is a real data limitation that requires continuous vigilance to minimize. However, we found that thorough examination of data reported as unknown or zero values was very helpful in revealing problems with specific variables and in improving data quality. We systematically queried States about data that included no unknown values when at least some unknowns could be expected. We learned much about both the data reported to us and how the data are collected from the field. For example, although we would expect unknown birth weight to be a rare event, we questioned States that reported no unknowns. In response to this query, one State found that reports of unknowns had been omitted because of a programming error.

The impact of missing data items was a major concern as a data quality issue at the NIMS Conference workshops for vital registrars and statisticians. One participant reported that her State had made an indepth study of the 0.5 percent infant deaths with missing birth weight. This State found that the largest portion were neonatal deaths occurring during the first day of life that would likely have been low birth weight infants. This bias could have created serious problems in analysis. Since more than 4 percent of all neonatal deaths in the NIMS project were reported in the unknown birth weight category, and the percent of unknown birth weight for all infant deaths among States ranged from 0 percent to 14 percent, this issue can be significant. Moreover, since 11 States reported at least as many deaths of neonates of unknown birth weight as of birth weights less than 500 g, the unknowns may compound the less than 500 g birth weight issue.

Pregnancy history variables are especially subject to ambiguous reporting of unknowns and zeros. Although the numbers of unknowns for these variables might have been small, we queried all States reporting no unknowns for numbers of either previous live births or other terminations (20 or more weeks' gestation). Thirteen States reported

no unknown number of previous live births, 2 additional States reported no infant deaths associated with 0 previous live births, and 11 States reported no unknown number of other terminations.

With respect to previous live births, the problems were manifold. The lack of zero values for one State was accounted for by the fact that parity values inadvertently had been submitted for previous live births. This error was easily corrected. Another State was unable to separate zero values from unknowns and therefore included zeros with unknowns. Among the 13 States reporting no unknowns for previous live births, several told us that unknowns or blanks, or both, are coded as zeros. Therefore, these unknowns or blanks are irretrievably zeros on their data tapes. Of unquantifiable and possibly greater concern was the response from some States that there were no unknowns in the data "because blanks are queried rigorously and not acceptable for these variables." We received this response both in telephone communications and at the NIMS Conference workshops. It became apparent that some States do not allow a certificate with an unknown value to be filed. Since it is extremely likely that some unknowns exist for at least a very small portion of pregnancy history variables, there is a strong possibility that there are some invalid zeros on the birth certificates. On such certificates, the data should be unknowns or should have been left as blanks that can be recoded after followup. Most of the invalid zeros are probably inadvertent errors, but they represent an important area where data quality could be significantly improved at the source.

At the Conference, some States reported having vital statistics field representatives who provide education to personnel at the facilities responsible for the completion or filing of certificates. However, there remains some confusion with the distinctions among unknowns, blanks, and zeros. Those who file certificates know that unknowns are frequently queried, that blanks are unacceptable, and that in fact "most of both of these" are actually zeros. Furthermore, they have learned that filing a zero answer, rather than a blank or an unknown, will avoid having the data queried. Additional effort at the source to explain the importance of correct information for these items, though admittedly resource intensive, may significantly improve the quality of many data items. It is important to deemphasize the "unacceptability" of blank and unknown responses; to reemphasize the distinct meaning of the unknown, blank, and zero responses; and to stress the importance of having correct information for programmatic and public health decisions.

'Although uniform definitions presumably are used, the attending physician decides which certificate to file, and the physician may take the easier route of filing one certificate (fetal death) rather than two (live birth and infant death) for infants who live only briefly.'

For the time being, zero values for parity are surely inflated. It is not possible to estimate the effect on the calculated mortality risks because a parallel distortion occurs on birth certificates of infants who die and who survive. There may be a bias toward more true unknowns among the zeros for early infant deaths than for survivors.

All the concerns mentioned apply equally to the other terminations (20 or more weeks' gestation).

In addition to the variables discussed, for the variable month prenatal care began, four States had no unknown values for either neonatal or postneonatal deaths. For number of prenatal care visits, two States reported no unknown values. Again, this is unexpected, especially since both the prenatal care variables had a significant number of unknowns among the 1980 births.

Analysis may reveal hidden data problems. Repeated use of a data set by researchers with different perspectives may reveal unexpected data problems even in much used and well-respected data bases. Several situations arose early in the NIMS analyses which illustrate that before relying on new findings, one must remain alert to the possibility of hidden problems in the data. For example, we discussed with one State its quite atypical gestational age distribution. Subsequently, the State found a problem in a nationally distributed computer program that it had been using extensively. A later version of the computer program truncates values more appropriately than the original program that the State had used for many years. As a consequence, during the time the State was using the original program much of its known

gestational age data went into the "unknown/other category." The State chose to re-do all the published annual reports for which it had used the obsolete program.

Another State posed a different challenge during early CDC analysis of NIMS data. In analyzing the risk of infant mortality on a State-by-State basis, we found that one State's low birth-weight-specific mortality appeared significantly lower than all others. We discovered that the birth weight distribution of this State's live birth data on the 1980 NCHS natality tape is very atypical for the low birth weight ranges and is not in agreement with the State's 1980 data. Therefore, we decided not to use NCHS live birth data for this State in any analyses using regional aggregates which include this State. Instead we redistributed the birth weight of this State's total live births based on the distribution of live births in all other States in the relevant census division (*1*).

We found a third problem when we analyzed State-by-State variations and racial differences in mortality risks in the "completed" NIMS Report data. When we explored one State's atypically low mortality risk for blacks, we found that one wrong programming step when the State produced its NIMS data had caused a misclassification of black infant race and substantial underreporting of black deaths for that State. Finding this well-embedded problem allowed us to revise all aggregate data for blacks in the preliminary report.

Summary and Recommendations

The NIMS project allowed us to compare data not accessible to individual States. In summary, we discovered problems when:

1. We reviewed States' 1980 birth certificate forms for compatibility with NCHS-NIMS definitions of variables—and we recommend alertness to differences between the U.S. Standard certificates of live birth, death, or fetal death and each State's forms for potential definitional peculiarities;

2. We compared birth weight distributions for each State—and we recommend exploration of any atypical findings both in examining individual State data and in future efforts to aggregate State data;

3. We compared distributions of maternal and infant characteristics for each State and queried variables for which there were unusually low or high unknown or zero values—and we strongly recommend continued suspicion of either unusual patterns or changes in patterns;

4. We learned that there is wide variation in the States' linking procedures and in the content and maintenance of linked files. This variation affects the States' infant mortality statistics. We recommend that State health officers maintain ongoing contact with neighboring States to keep the interstate transcript exchange functioning well;

5. We explored all unusual value patterns reported for unknown and zero categories—and we recommend (*a*) continued alertness to the pervasive problems involved in the reporting of these values and (*b*) continued efforts to improve instruction on the distinct meanings and importance of accuracy of these values. Such instruction should be repeated often enough to take into account the high turnover rate of personnel completing and filing certificates;

6. We compared States' infant mortality risks to the national, U.S. census region and division risks—and we recommend that all large deviations be explored for previously undiscovered data problems of the type discussed in this paper.

We have presented some of the major data quality issues related to the use of linked birth and infant death certificates. The information obtained from the NIMS project is providing researchers with new insights into the problems related to infant mortality.

References

1. Centers for Disease Control: National Infant Mortality Surveillance report, 1980. Atlanta, GA. In press.
2. Hogue, C. J. R., et al.: Overview of the National Infant Mortality Surveillance (NIMS) project—design, methods, results. Public Health Rep 102: 126–138, March-April 1987.
3. Lambert, D. L., and Strauss, L. T.: Analysis of unlinked infant death certificates from the NIMS project. Public Health Rep 102: 200–204, March-April 1987.
4. SAS User's Guide: Basics, 1982 Edition. SAS Institute, Inc., Cary, NC, 1982.
5. Association for Vital Records and Health Statistics: Agreement for administering the Interstate Vital Records Exchange System, adopted Kansas, MO, 1982.
6. Prager, K., Flinchum, G. A., and Johnson, D. P.: The NCHS pilot project to link birth and infant death records: stage 1. Public Health Rep 102: 216–223, March-April 1987.