NEW APPROACHES TO CODING AND ANALYZING MORTALITY DATA

Robert A. Israel
Marvin C. Templeton
Marshall C. Evans
National Center for Health Statistics

During the past seventy-two years of collecting national mortality data from the vital registration system in the United States, the coding of the demographic and medical data has been a necessary clerical chore. In addition, the analytic possibilities involving cause of death have been limited by the principle of ascribing one and only one cause to each death, as proscribed by international agreement, thus resulting in a loss of the other medical information which may appear on the death certificate. This paper presents a discussion of computer systems that have been developed, and some that are in the process of being developed by the National Center for Health Statistics, that classify and code demographic and medical data items from death certificates by computer processes and as a by-product yield additional cause of death information which may be utilized for new and more extensive analyses.

One of the most complicated kinds of statistical data classification has been the assignment and coding of the underlying cause of death. This concept involves the determination of "the single disease or injury which initiated the train of morbid events leading directly to death or the circumstances of the accident or violence which produced the fatal injury." 1/

The World Health Organization has provided the statistical classification structure and the attendant rules for determining the underlying cause of death through the decennially revised International Classification of Diseases. 1/ While this classification system is comprehensive and invaluable from many points of view, it leaves for the user a number of serious problems. Some of these difficulties are: (1) the long period of training required to develop a competent nosologist schooled in the classification and coding of mortality medical data and the related problem of staff retention in an area of work that requires exactness and controlled productivity; (2) the difficulty in achieving desired levels of data comparability and accuracy brought about by complicated rules and classification structures, inexactness in reporting of some medical conditions, and human limitations in the broad range of knowledge required by a nosologist; and (3) the selection of a single underlying cause of death resulting in the loss of the remaining information reported on the medical certification portion of the death certificate. Much of the lost information may be relevant to current health problems.

To meet the expanding needs for greater utilization of mortality medical data and to take steps which would overcome some of the problems encountered in the traditional coding procedures, the National Center for Health Statistics has developed an automated computer system which provides statistical data on all medical information reported on death certificates. The system is designed to generate the underlying cause of death for the first time through computer methodology by applying the international coding rules to each medical relationship given by the certifier of the death in much the same manner employed by nosologists in manual operations. The system carries the name ACME which is the acronym for "Automated Classification of Medical Entities."

The system requires the entry onto computer tape of all diseases, conditions, accidents, and injuries given in the medical certifier's statement. The information is recorded onto tape utilizing codes based on the Eighth Revision International Classification of Diseases Adapted for Use in the United States (ICDA). Special instructions are used in applying the ICDA codes to all diseases, conditions, and injuries. Once the medical certification is placed on magnetic tape in coded form and in the same sequence as they appeared on the death certificate, the medical relationship of the entries are matched against a series of decision tables which relate all ICDA codes to one another in terms of each international rule. Defining the specific content of each international rule decision table presented one of the most difficult systems development problems in that it required a deliberate decision on the relationship of each detailed code to every other ICDA code relative to the purpose of each international rule. Traditionally, such decisions have been made by individual coders who have had limited guidelines on the causal relationship of diseases and on priorities of some conditions over others.

The computer program is written in PL1 Computer Programming Language. It provides for storage of all decision tables in core storage, minimizing reference time requirements, which, in turn, increases output speed. All functions are performed and the selection of underlying causes is accomplished at a rate of approximately seven records per second.

The system was developed, tested, and implemented on the NCHS IBM-360 Model 50 computer which had 256K core storage until the system's capacity was recently expanded. Storage requirements dictate that the system, in its present form, cannot be accommodated on a computer of less capacity.

Decision table content is updated by introducing individual changes when made by the staff. The tables are routinely maintained on disc storage and a special program provides for insertion or deletion of content coupled with a print-out of the updated tables with special notations of the changes for visual verification.

Approximately 5.5% of the data records are rejected for manual processing and are printed out in full detail with appropriate messages defining the reason, a means of handling the exceptional

cases not conforming to the provisions of the system. In addition, whenever a disagreement is encountered in the quality control process, and there is manual coding of a sample of records serving as a quality control, then a full description of the detailed steps the computer system applied is printed out including each decision table reference and the order in which the decisions took place. This leaves no question as to how the computer system derived its final answer. This is one means by which adjustments to the decision tables are identified, defined, and incorporated into the system. Furthermore, if a particular decision is changed and is considered sufficiently significant to warrant correction on records previously processed, it can be accomplished by simply resubmitting the data file to the system based on corrected decision tables.

Manual application of the International Coding Rules is subject to varying degrees of inconsistency. The computer system provides an avenue of absolute consistency and a means of isolating troublesome certifications for further study. An important by-product is the full documentation for the first time of assumptions and decisions going into cause-of-death classification. Heretofore, documentation has been given in terms of guidelines with minimum reference as to how they are to be exercised in specific situations. Automating this element of the classification process gives users detailed insight into the data and permits more intelligent analyses of the end product.

The same data that serve as input to the ACME system for the assignment of the underlying cause of death, i.e., the ICDA codes for each diagnostic term appearing on the medical certification of death, can also serve as the data inputs to multiple cause of death analyses which would take cognizance of all conditions reported at death on the certificate. This approach is not in itself new but at the national level prior to 1968 the coding of more than just a single underlying cause of death has been undertaken only five times since 1900 - in 1917, 1925, 1936, 1940, and 1955. The underlying (or principal) cause of death and one associated cause were coded in 1917, 1925, 1936 and 1940; in 1955 all reported information was coded. A single table showing the cross tabulation of underlying and contributory causes was published without comment for the data years 1917, 1925 and 1940 in the annual vital statistics publications of the United States for the years 1918, 1925, and 1940 respectively. 2-4/ A paper presented to the American Public Health Association in 1923 presented a brief analysis of the 1917 data and strongly recommended additional work on multiple causes of death. 5/

Continued interest in multiple-cause tabulations was stimulated by the Fourth International Conference for the Revision of the International List of Causes of Death. International comparisons of the procedures for selection of the primary cause of death indicated that comparability of death rates could not be achieved on an international basis until there was more knowledge of the contributory causes of death. An extensive study of multiple causes of death was then undertaken for 1936. Two condensed reports arising from these data were published in 1939 and 1940 but they did not contain the full set of tables. 6,7/ Associated causes of death were again coded for 1940 and a table was included in the regular annual vital statistics volume for that year. 4/

During the early 1950's a number of activities both at the national and at the State level, were undertaken in connection with multiple-cause-of-death studies. By 1956, the National Office of Vital Statistics (now the Division of Vital Statistics of the National Center for Health Statistics) had developed a manual of instructions for the coding of multiple causes of death and work proceeded slowly in the coding of a sample of 1955 death records. The results of this activity were published in various journals and in a Supplement to the 1955 edition of Vital Statistics of the United States. 8-11/ This latter publication, however, did not appear until 1965. In addition, at the international level, two meetings - one in 1967 in London and one in 1969 in Geneva - brought together a number of interested nations for the purpose of exploring uses of multiple cause analysis and methodology but with heavy emphasis on minimum standards for international comparability.

For all of the activity relating to multiple cause analysis dating back to the early part of the century and continuing sporadically to the present, it might be assumed that the major problems have been solved, but this is not the case. There still remain a number of key issues.

First of all is the question of the appropriateness of the medical diagnoses appearing on the death certificate. Are they accurate? This question, of course, is applicable to not only multiple cause of death analysis but to the more conventional underlying cause of death as well. With all the attention focused on mortality statistics over the years, it is necessary to continually be concerned with this question. Studies in the past have indicated some problems in accuracy, but nowhere near enough evidence has been found to invalidate the usefulness of mortality data. However, not enough is known about this problem and more effort must be directed at studies of accuracy of medical certification on death certificates. Closely related to the problem of accuracy is the question of completeness of reporting of conditions. If attention is now to be focused on all of the conditions listed on the death certificate, what conditions should be listed? Over the years, educational efforts have been made to elicit from the certifying physician enough information to satisfactorily indicate the single primary or underlying cause to which the death is to be assigned. Any conditions listed on the certificate other than the underlying cause hopefully were useful in arriving at that assignment, but no other use was made of additional medical data. However, the multiple cause approach changes all of that. Now there is interest in each condition on the death record. How many conditions do physicians list? How many should they list? For underlying cause purposes a certificate which simply indicates "pulmonary tuberculosis" is statistically equivalent to one which indicates that the immediate

cause of death was "congestive heart failure" which was due to "bronchiectasis" which in turn was due to "pulmonary tuberculosis". For multiple cause purposes there is an obvious difference in the amount of information available and yet the two cases may have been medically very similar. How uniformly do certifying physicians include conditions which contribute to death but are not related to the immediate cause? What about conditions present at death but unrelated to the immediate or underlying cause? Obviously a new educational effort will have to be made if any degree of consistency of reporting is desired. On the other hand, we are not ready to abandon the underlying cause concept. The multiple cause data are considered a useful supplement but not a replacement. Therefore any change in instructions will have to be carefully phrased so as not to upset the one approach for the sake of the other. Related to these problems is the format of the internationally recommended medical certification. Should this format be revised to better elicit the kinds of information desired? It is hoped that experimentation in this area leading to recommendations to the World Health Organization will be undertaken in various parts of the country and around the world.

The second major area of concern lies with the structure and content of the International Classification of Diseases (ICD). How should this important statistical tool be modified to accommodate multiple condition analysis? Presently, there are rubrics in the ICD which are already combination categories intended to provide for the joint reporting of certain combinations of diseases such as category 404, Hypertensive heart and renal disease. For multiple cause purposes, does this title represent one condition, two conditions, or three or more conditions? A closely related problem is the coding and counting of conditions such as hypertensive cardiovascular arteriosclerosis (one term with one code number in ICD) as opposed to the reporting of cardiovascular disease due to arteriosclerosis due to hypertension reported on three different lines of the death certificate. Should these two situations be handled differently or the same for statistical purposes? What kinds of coding rules should be developed to code multiple conditions on death records? There are numerous problems of this type which arise because of the structure and intended purpose of the ICD. What recommendations should be made to alleviate some of these problems? The third area of concern lies with the kinds of output or statistics that result from this approach. What kinds of tabulations and analyses are envisioned? What uses can be made of such data? Here, as with the other areas of concern, there is not complete agreement. However, some very basic tabulations have been generally agreed upon among the interested countries. These basic tabulations will for example, show for a selected list of causes of death the relationship between the underlying cause and the associated causes reported with the underlying cause, the number of deaths where each selected condition is mentioned whether or not the condition is the underlying cause, and the total number of times the selected conditions are mentioned on death certificates. It should be noted that the number of deaths involving a given condition and the number of times that the given condition is mentioned on death certificates are not necessarily the same. Other basic tabulations might present data on combinations of conditions that appear more or less often than might be expected if the conditions are independent of each other.

There have been a number of questions or problem areas raised and not many, if any, answers have been supplied. These questions are by no means exhaustive; it is hoped that they are sufficiently suggestive of the kinds of problems being faced and solutions that are being sought. This does not mean that the National Center for Health Statistics does not have any approaches to the technique of multiple cause analysis. On the other hand, ideas and suggestions from users or potential users of these data are encouraged. It is important to ask once again the question raised by Dorn and Moriyama: "Why do we want statistics on causes of death?" 8/

These authors stated that the uses to which statistical data are put are basically determined by the available data. They further stated that the types of information one might reasonably expect cause-of-death statistics to provide in order to maximize their usefulness are as follows:

a. accurate reflection of the conditions contributing to the fatal outcome in the opinion of the medical certifier.

b. the relative importance of the various diseases, injuries and acts of violence as causes of death and

c. reliable representation of the time trend of the frequency with which the various conditions are reported as bringing about death.

It is believed that multiple cause of death data will maximize the use of available diagnostic information. It will be possible to make a count of all reported conditions as well as an unduplicated count of deaths. There will be a basis for analyzing mortality trends for various diseases which is not now possible because of the loss of information resulting from the selection of a single condition as the underlying cause of death. Multiple cause tabulations will provide a great deal more data than are now available. They will avoid some of the long standing objections to the underlying cause concept. However, if multiple cause data are to realize their full potential and if the National Center for Health Statistics program of routine publication of such data is to be of value, then significant interaction between the producers and the consumers of the data must come about. For the immediate future, decisions are being made in order to produce data that can be reviewed and evaluated. In the long run, improvements in the basic input as well as in the output and analysis will hopefully add significantly to the vital and health statistics field.

Thus, the ACME system presents several benefits in the coding and analysis of mortality data. It provides for the automated selection of the underlying cause of death in a uniform manner, eliminating to a very large extent the intercoder variations inherent in a complex manual system. It provides, through the same basic input codes, the ability to tabulate and analyze more than one cause per death. In addition, the system gives baseline information for evaluating the International Classification and the rules for the selection of the underlying cause, an invaluable set of data useful for the periodic revision of the ICD.

These benefits are derived with savings in training and overall manpower requirements. Previous attempts to code multiple causes as well as underlying causes have proved to be very costly. Coding of multiple causes requires manpower resources equivalent to that used in coding underlying causes alone. Manpower for detailed coding required by the ACME system is about one-fourth greater than that required to manually select the underlying cause. The contrast is attributable to simplification of the manual coding process. Training requirements for nosologists have been reduced from 12 to 18 months to from 3 to 6 months.

It should be noted that the ACME system has been introduced by our colleagues in the Vital Statistics Section of Statistics Canada and we jointly believe that it is a significant achievement in keeping vital statistics data handling and processing apace with advances in methodology and technology.

There are other challenging research projects in the data preparation and processing field under long range development in the National Center for Health Statistics.

One project underway is the development of a computer system named CONTEXT which will accept full text or standard abbreviations of the medical terms encountered on death certificates and will convert these terms automatically into their corresponding ICD code numbers. Such code numbers then could form the input to the ACME system, thereby eliminating the manual assignment of ICD codes for each condition as is now the practice.

Another such project has as its goal the automation and standardization of non-medical or demographic data items on vital records. The objectives of this project are:

1. To automate to the highest practical level the entire spectrum of data collection, handling, and processing of the nation's vital statistics data.

2. To develop standardized terminology, definitions, and data classification detail for vital statistics applications at various levels of government.

3. To provide a single data handling system which will meet the needs of Federal, State, and local vital statistics programs.

For conversational conveniences, we have dubbed this project ASSIST, the acronym for "Automated Standardized System for Indexing and Statistical Tabulations."

The state of the art in data processing has developed to a degree that technologically there should be no serious barriers in meeting the objectives of ASSIST. Data needs are clear, which leaves to human resolve and commitment the matter of developing standard terminology, definitions and classification detail.

The rising costs of manpower and equipment and the growing gap between vital statistics methods and available technology bring considerable pressure for upgrading our procedures through efforts such as ACME, CONTEXT and ASSIST. These new approaches to coding and analyzing mortality data will provide opportunities for the production of more uniform, standardized statistics on a more timely basis, and will eliminate much of existing duplication of effort between the various levels of government, while at the same time provide for sufficient flexibility in the system to meet the needs of federal, State, and local health statistics programs.

REFERENCES

1. Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death. Volume 1, World Health Organization. Geneva, Switzerland, 1967.

2. U.S. Bureau of the Census, Mortality Statistics, 1918, U.S. Government Printing Office, Washington, D.C., 1920

3. U.S. Bureau of the Census, Mortality Statistics, 1925, Part I, U.S. Government Printing Office, Washington, D.C., 1927.

4. Vital Statistics of the United States, 1940, Part I, U.S. Government Printing Office, Washington, D.C., 1943.

5. Dublin, Louis I., and Van Buren, George H., "Contributory Causes of Death--Their Importance and Suggestions for Their Classification," American Journal of Public Health, Vol. 14, No. 2, February 1924, pp. 100-105.

6. U.S. Bureau of the Census, "Deaths from Alcoholicm, United States: 1936," Vital Statistics - Special Reports, Vol. 7, No. 59, 1936.

7. Janssen, Theodore A., "Importance of Tabulating Multiple Causes of Death," American Journal of Public Health, Vol. 30, No. 8, August 1940, pp. 871-879.

8. Dorn, Harold F., and Moriyama, Iwao M., "Uses and Significance of Multiple Cause Tabulations for Mortality Statistics," American Journal of Public Health, Vol. 54, No. 3, March 1964, pp. 400-406.

9. Moriyama, Iwao M., "Chronic Respiratory Disease Mortality in the United States,"

Public Health Reports, Vol. 78, No. 9, September 1963, pp. 743-748.

10. National Cancer Institute, "New Numerators for Old Denominators--Multiple Causes of Death," by D. E. Krueger, National Cancer Institute Monograph, No. 19, Washington, D.C., 1966, pp. 431-443.

11. Vital Statistics of the United States, 1955, Supplement, U.S. Government Printing Office, Washington, D.C., 1965