Records and Statistics

R. ZEMACH, PhD, and T. R. ERVIN, MA

Dr. Zemach is associate professor, department of biostatistics, University of Michigan School of Public Health, Ann Arbor, and Mr. Ervin is chief, bureau of administrative services, Michigan State Department of Public Health. Tearsheet requests to T. R. Ervin, Michigan State Department of Public Health, 3500 N. Logan, Lansing, Mich. 48914.

THE nature and role of statistical information systems are receiving considerable attention at the present time, especially at the Federal level. Although impressive progress has been made in the collection, processing, and use of statistics, there is only a beginning consensus on what constitutes an optimal program for the future. Congress has supported a new initiative in a Federal-State-local cooperative health statistics system. However, the $1.8 million appropriation for initial research and development in 1971–72 was made with the expressed concern of members of Congress over the magnitude of future requirements for funds and over the intention and capability of the health system to use efficiently the data that are produced (*1*).

The need for restraints, as well as progress in developing an effective and economical health information base, was also reflected in the appointment of a committee in 1971 to evaluate the National Center for Health Statistics and related health data systems in the Department of Health, Education, and Welfare.

An inquiry carried out among major producers and users of health statistics has focused renewed attention upon issues such as: What is the relationship between general-purpose or baseline statistics and program data? What are the characteristics of record systems when compared with statistical data collections? Do the data needs for health services planning, evaluation, research, and training require data collection completely separate from management information systems? Are program-generated data too biased by the nature of the program to be used as baseline information? What level of compatibility can be achieved

among baseline data sets and program-generated data short of record linkage based upon a universal numbering system?

In this paper, we focus on the statistical (quantitative) information needed for the production and utilization of health services, but most of the ideas expressed may be applied equally well to other areas of public welfare. Not all the problems mentioned can be discussed in detail, but we suggest that a core issue, cutting across such questions, is the need to recognize the differences, similarities, and relationships between recordkeeping and statistical data collection. An understanding of the interplay and resulting configuration has implications for the future productivity of information systems, especially the question of how and where funds should be allocated for further system development.

Recordkeeping and Statistical Data

The term "data item" is defined as a collection of measurements or qualifications of a single factor or attribute for some population. For example, a recording of age, income, race, and disability state of the population of a neighborhood would constitute a set of four data items on that population.

In recordkeeping, the major requirement for a data item is that each individual datum be retrievable in its explicit relationship to the person or unit from which it was drawn. Records must contain unique identifying information; otherwise, the recordkeeping system cannot fulfill its function. A record is required for each person or unit in the client population for whom the system is maintained or, conversely, the client population consists exactly of those persons or units for whom there are records. Records are not substitutable, and if an item is missing from a record, the record may not fulfill its function for some person or unit.

Computational capability is not needed. In a computerized recordkeeping system, computer technology is used to enhance acquisition, communication, and retrieval, rather than computation as such. Speed and ease of retrieval are essential. In some situations, recordkeeping is required by law, and thus the storage of records is as important a function as retrieval.

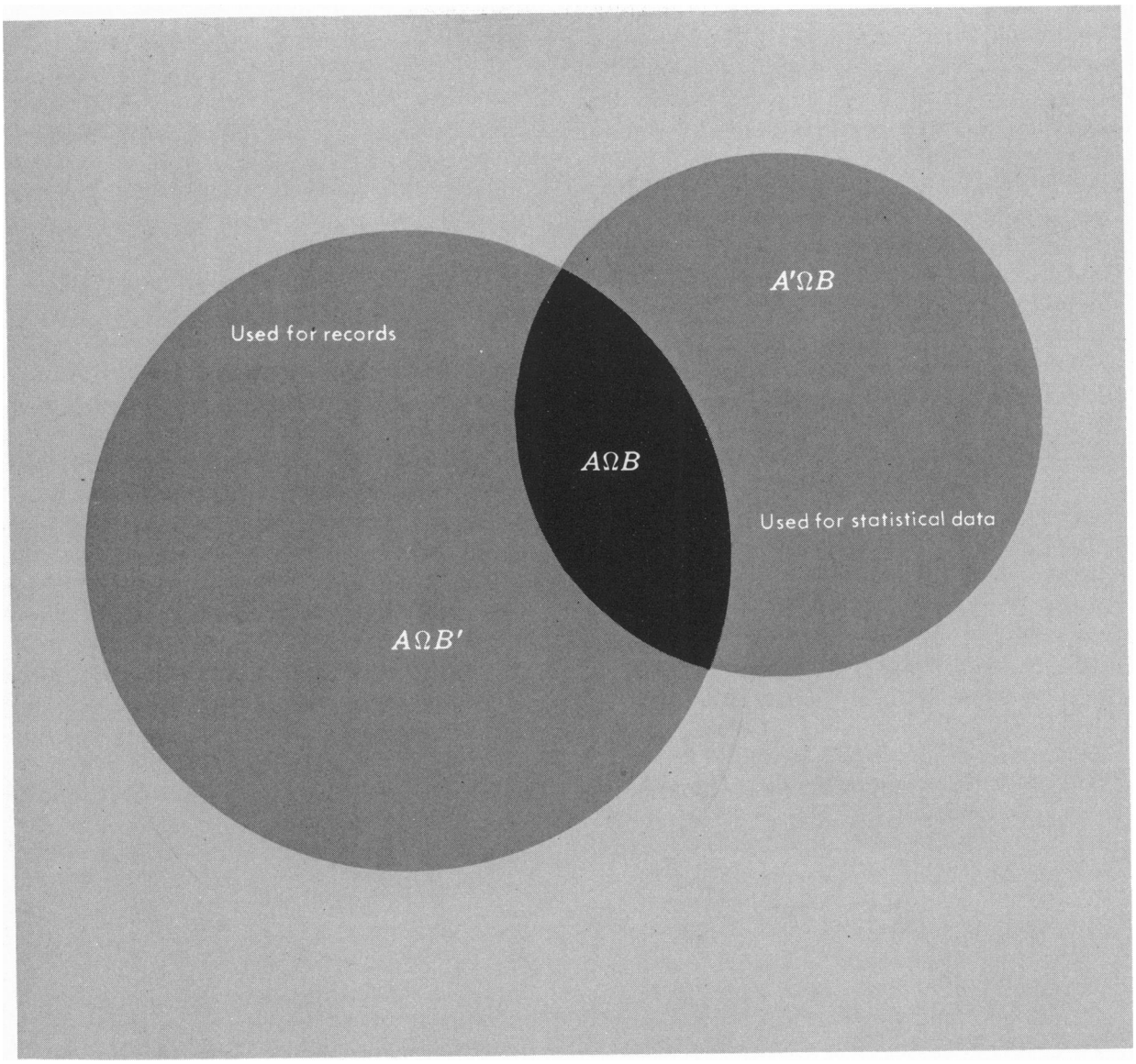
Examples of recordkeeping systems are clinical files used for patient care, Medicaid patient and provider records, licensure lists, and registries of vital records.

In a statistical data system, the data are of interest only when produced in aggregated or summary form. Unique identification of a datum with a person or unit is irrelevant; in fact, the records need not have any individual identification. Under appropriate circumstances, information based on data in the system can be extrapolated to a larger universe that includes persons or units whose records are not in the system. That is, sample data will suffice for the purposes of the system. Data items (sets of measurements) rather than individual measurements will be retrieved. In a computerized system, the computers are needed to compute. Storage of data is a secondary consideration; if the data are not to be retrieved, there is no basis for storing them.

Clearly a recordkeeping system and a statistical data system have considerably different characteristics with respect to purpose, management, computer requirements, and quality control considerations. A statistical data system can tolerate a certain number of missing data or errors, and still produce all the information required from it. Missing data or errors in a record system, however, may seriously affect its function.

Consideration of data systems in the real world, rather than in concept, quickly establishes that many of these systems do in fact serve for both recordkeeping and statistical data production. A set-theoretic configuration may be used to show the relationship between the two concepts. Consider a hypothetical universe, as shown in the chart, of all possible data items as previously defined. The set A comprises data items that are used in recordkeeping, the set B data items that are used to produce statistical data, and the set $A \Omega B$ data items used for both recordkeeping and statistical data. If a data item is in A but not in B (in $A \Omega B'$), it is used as part of the records of individual persons or units, but not used, or not usable, in aggregated form as statistical information. If a data item is in B but not in A (in $A' \Omega B$), its elements are not intended to be used in relation to the individual persons or units on which they are based, but only to produce statistical data. And, of course, outside of $A \Omega B$ is a world of uncodified, uncollected, and perhaps, unnamed data.

These sets define neither the systems nor the records that are in the systems but classify individual data items according to their utilization within a system. A single system may have some items used only as a part of records (for example, name and street address), some items used both



as part of individual records and to produce statistical information (for example, age and income), and possibly an item used only for statistical purposes (although if it exists as part of an individual record, it will tend to be retrieved as such).

The characterization of statistical information as baseline data is a reference to the extent of population coverage of the data items being used to produce the information relative to the question being asked, regardless of whether the system from which the data are drawn is a recordkeeping system or not. As a ready illustration, in the vital records system the linkage of infant death and live birth data can provide national baseline data on infant mortality (*2*). If all citizens age 65 and older receive their health care through a Federal program, the program records can potentially produce baseline information on one segment of the population. Whether or not special statistical sampling data can be regarded as baseline or not similarly depends on the extent of the population sampled.

Where Are All the Data?

Most of the data collected are in set *A;* that is, they are stored in recordkeeping systems, although many of the items are also used to produce statistical data. Therefore, most statistical data which are or could be available to us must be produced

from recordkeeping systems. A relatively small percentage of items are obtained specifically for statistical purposes; these include data from sample surveys. Further, the data systems that are used as both recordkeeping systems and statistical data systems are almost always established with recordkeeping primarily in mind.

The best known example includes the vital registries, now also known as vital statistics systems. These registries were established by law to register the occurrence of a vital event. They were also seen as depositories for the records, so that a person might retrieve a birth certificate or other record as evidence of the occurrence. Vital records systems originated for recordkeeping purposes despite the fact that the collection of statistical data on mortality and births was a springboard for the development of the science of statistics (*3*).

In States and localities, these systems, by whatever name they are known, are used as both recordkeeping systems and as statistical data systems. At the Federal level, the vital records system is a statistical system, beause it is not the function of the National Center for Health Statistics to retrieve a person's birth or death certificate. Therefore, the Center's purposes are served through sampling of records, although all the records amassed by the States and localities are currently reproduced in its files.

Too Much But Not Enough

Two complaints which are often expressed are "We don't have any data" and "We have to keep too many records." These complaints seem paradoxical: if so much information is being recorded, why can we not find out more about what is going on? Throughout the vast complex of production and utilization of health care, there are many data systems. Except for a relatively few, their contents are defined by some institution or some program for health services, and they are pieced together as recordkeeping systems. They often include records of persons who receive a service, the service provided, the providers, the amount and type of resources used, and the costs. To varying degrees, these record systems also provide some statistical data, if only because reporting is required by law or needed for managerial planning and evaluation. But when a system is conceived as a recordkeeping system, with secondary consideration given to producing limited statistical data, it is ill equipped to serve any broader purpose in the total system of health care. Therefore, we have the phenomenon of other data units being assembled, through surveys for example, to provide some statistical information, when potentially comparable data units are locked away in record systems.

We see an example of "too much but not enough" in the collection of ambulatory care data. The bulk of all health care, in terms of incidents of service provided, is given in an ambulatory setting. Often it is said that we know little of what goes on in private medical and dental practice, meaning we have few aggregate statistics. Yet the provider typically collects a number of data items concerning his patients and has fairly complete records on encounters. In only a few large ambulatory care organizations do physicians and dentists have these records stored in such a way that statistical data are easily retrievable (*4*).

A second major depository of ambulatory care data is the collection of record systems of categorical health programs. These programs serve large segments of our population, particularly the poor, and each record system contains a wealth of detail on the types and costs of health services provided and on the health of people (*5*). No one would minimize the difficulties in getting adequate statistics for planning, evaluation, research, and training, as well as for management, from such health record systems. These functions are carried out at different levels and from different perspectives, so there are differences in the kinds of data required, the level of detail and timeliness, and the extent of compatibility required to achieve a designated purpose. Program records often provide fragmented population coverage, and record items are not standardized from one program to another. Even with the use of computers, it may be inefficient to retrieve statistical data when a system is designed and managed as a recordkeeping system.

It is easy to conclude that although we are overwhelmed with recordkeeping, the paucity of usable and useful statistical data will require an equally overwhelming venture into statistical data collection. Two considerations, however, show the need for careful reexamination of this conclusion: (*a*) a consideration of the potential funding of data-collection activities and (*b*) a consideration of the costs of special statistical data collection compared with the benefits derived. Although "better data" are frequently demanded in support of cost-benefit analysis, the cost-benefit ratio concept is

seldom applied to the production of statistical data as such.

Where Does the Money Go?

Money for records is usually included as part of a program or institution budget. This money is taken for granted. No one questions the need for recordkeeping funds when a program is initiated, and furthermore, funds will continue to be allocated for recordkeeping as long as the program exists (*6*). A statistical data collection, on the other hand, is often viewed with some skepticism, must provide explicit proof of its usefulness, and faces a continuing threat of loss of funds (*1*).

Each federally funded program requires the establishment of a recordkeeping system, and there are more persons in the Health Services and Mental Health Administration working with program statistical systems than there are working in the statistical functions of the National Center for Health Statistics (*7*). The funds appropriated in 1971 for research and development of a Federal-State-local cooperative health statistics system ($1.8 million) are less than the funds available for development of recordkeeping in a single program such as Medicaid. The contrast is even greater in nongovernmental health activities; seldom are funds used for the collection of statistical data apart from recordkeeping unless the data are needed for Government grant applications. Large sums are being spent to improve the efficiency of health care through the use of computer technology, but research and development focuses on acquisition, storage, retrieval, and communication of records, with production of statistical data too often ignored.

Inasmuch as this disparity between what is provided for management record systems and what is provided for statistical systems per se has prevailed strongly in the past and can be expected to do so in the future, we question whether development of purely statistical systems could lead us to a solution to the problem of lack of information. Instead, the potentials in the development of record systems to serve both management and statistical interests should be recognized. With standardization of contents, definitions, reporting periods, and other features of recordkeeping, data from various record systems could be linked according to geographic, demographic, economic, or health-related variables included in the records, to provide needed summary data. The records of individual persons need not be linked, except as required for management of service delivery.

Which Data Get Used?

Field experience with surveys which have been planned and operated apart from program operations suggests that the commitment to use statistics may vary directly with the involvement of the users in the statistical collection (*8*). We question the usefulness of statistical systems developed apart from some programmatic emphasis. If program planners and managers perceive statistical data as something produced by "those statisticians," they may fail to see any relationship to their own program needs, and the data will have little impact on health care development.

Similarly, statistical use of a recordkeeping system needs careful planning by statisticians and program staff working together, so that the information produced is no more and no less than what is needed.

Where Do We Go Now?

A consideration of where the money goes and is likely to go in the future and a consideration of utilization of data lead to the conclusion that cost-benefit ratios can be optimized by the development of the statistical data production capabilities of recordkeeping systems. True, there will always be information that can be obtained only by special surveys or collections. But because of the cost of such collection, the need for the information and the lack of alternative sources should be carefully established. Collaboration of statisticians in the development of record systems could, in the future, lead to limiting and sharpening the focus of separate statistical data collections because recordkeeping systems could be exploited for relatively moderate costs to produce much of the information needed. Coherence in the overall system will require much greater attention to establishment of standards, definitions, and units of measurement, co-terminous periods for data reporting, flexibility in data retrieval, and capability for the sampling of records. At present, even within federally funded programs, variations in definitions and classifications rule out statistical compatibility to a significant extent (*5*).

An important and related issue is the timeliness of information. The recordkeeping system used to manage patients or operate a program or institution must be kept up to date. Even our largest industries are expected to file accurate detailed reports within a few months following the close of their fiscal year for income tax purposes. Yet, as the editor of Science magazine pointed out

recently, the latest government statistics are often several years old (*9*). If management-level industrial recordkeeping systems are capable of such a high degree of timeliness, then a national health statistics program might well research, develop, and take advantage of potential reporting capabilities in health care programs and institutions to produce indices of health service production, use and costs, and of health status on just as timely a basis.

Compatibility and linkage remain major problems. When the records are viewed in terms of statistical use, standardization of classifications and definitions across record systems is a rudimentary essential not yet realized. Varying degrees of linkage are required for functions ranging from utilization studies, in which linkage to denominator data on population at risk is paramount, to clinical and epidemiologic studies, which may require sophisticated linkage to other systems. Reservations related to confidentiality continue to preclude a universal numbering system in the United States; however, this question does not impede the development of greater compatibility among summary data from various record systems. Such standardization can substantially improve the utility of records for quality statistics.

The National Center for Health Statistics has charged the State centers to ". . . concentrate . . . data collection toward general purpose statistics, as distinguished from specific program statistics" (*10*). This should be viewed as a charge to develop a structural interdependence between program staff and statisticians, leading toward more useful record systems, without assuming that general-purpose statistics require new data collection completely apart from program records.

An example of the utilization of multiple record systems to produce statistical information is the Family Planning Program record system in Michigan. Uniform record systems were organized (and funded) to provide management data for clinics. Under the guidance of the State health department's Center for Health Statistics, however, the record was designed to provide statistical information on the overall program at the State level and to provide the Federal Government with designated data in machine-readable form. The data available from the Family Planning Program's records can support rigorous statistical study for evaluation purposes. An effective direction for the future would be the incorporation of compatible data items in the records of other publicly funded programs providing family planning services, and eventually, the production of comparable statistical data by the private sector.

In this discussion of the implications of using record systems for statistical data, we do not imply that existing statistical systems should be discarded, nor do we mean to discourage current and future work to develop such systems. But, it appears that the first step toward comprehensive and timely statistics is to insure that we have developed and are making optimum use of what is or can be made available from records. The next steps toward statistical systems can then be taken on firmer ground.

REFERENCES

(*1*) U.S. House of Representatives, Committee on Appropriations: Departments of Labor and Health, Education, and Welfare appropriations for 1972. Appropriation hearings on health services research and development system. Pt. 2, 1st sess., 92d Cong., U.S. Government Printing Office, Washington, D.C., Apr. 23, 1971, pp. 255, 256.

(*2*) Chase, H. C.: A study of infant mortality from linked records: registration aspects. Am J Public Health 60: 2181–2195, November 1970.

(*3*) Graunt, J.: Foundations of vital statistics. *In* The world of mathematics, edited by J. R. Newman. Simon & Schuster, Inc., New York, 1956, vol. 3, pp. 1421–1435.

(*4*) National Center for Health Services Research and Development: University medical care programs: evaluation. Conference proceedings. DHEW Publication No. (HSM) 72–3010. U.S. Government Printing Office, Washington, D.C., December 1971.

(*5*) Association of State and Territorial Health Officers: A staff guide to health data systems review. Health Program Reporting System Project staff paper, Washington, D.C., Feb. 14, 1972.

(*6*) Nitzberg, D. M.: The basic neighborhood health center data system. Am J Public Health 61: 2065–2084, October 1971.

(*7*) National Center for Health Statistics: The mission and policies of the National Center for Health Statistics. DHEW Publication No. (HSM) 72–1201. U.S. Government Printing Office, Washington, D.C., April 1971, p. 6.

(*8*) Michigan Department of Public Health: Evaluation, Michigan Health Survey, 1967–71. Lansing, Mich., December 1970.

(*9*) Abelson, P. H.: Federal statistics. [Editorial.] Science 175: 1315, Mar. 24, 1972.

(*10*) National Center for Health Statistics: A State center for health statistics. An aid in planning comprehensive health statistics. *In* Conference series. Public Health Conference on Records and Statistics. Doc. 626. U.S. Government Printing Office, Washington, D.C., 1969, p. 13.