Methodological, Technical, and Ethical Issues of a Computerized Data System

Cindy A. Rice, MSPH, Michael A. Godkin, PhD, and Robin J.O. Catlin, MD
Worcester, Massachusetts

This report examines some methodological, technical, and ethical issues which need to be addressed in designing and implementing a valid and reliable computerized clinical data base. The report focuses on the data collection system used by four residency based family health centers, affiliated with the University of Massachusetts Medical Center. It is suggested that data reliability and validity can be maximized by: (1) standardizing encounter forms at affiliated health centers to eliminate recording biases and ensure data comparability; (2) using forms with a diagnosis checklist to reduce coding errors and increase the number of diagnoses recorded per encounter; (3) developing uniform diagnostic criteria; (4) identifying sources of error, including discrepancies of clinical data as recorded in medical records, encounter forms, and the computer; and (5) improving provider cooperation in recording data by distributing data summaries which reinforce the data's applicability to service provision. Potential applications of the data for research purposes are restricted by personnel and computer costs, confidentiality considerations, programming related issues, and, most importantly, health center priorities, largely focused on patient care, not research.

Recent years have seen an increased acceptance of family practice as an academic discipline and a growing interest in research within this relatively new medical field. The literature is beginning to emerge with articles describing the "Family Practice Experience," especially through the use of computerized data collection systems which gather varied information about patient encounters, diagnoses, and management.[1-7]

The emphasis of such articles has been primarily on the results of research conducted from computerized data bases. Often neglected, however, are important methodological issues which relate to the implementation of a computerized data collection system.* Such issues impinge upon the data's reliability and validity, which must be evaluated before research findings can be applied in any meaningful way. This paper addresses some of the important methodological issues, both techni-

From the Department of Family and Community Medicine, University of Massachusetts Medical Center, Worcester, Massachusetts. Requests for reprints should be addressed to Ms. Cindy A. Rice, Department of Family and Community Medicine, University of Massachusetts Medical Center, 55 Lake Avenue North, Worcester, MA 01605.

*One exception to this has been a discussion by Levinson (1978)[8] about health care and information management, in which he addresses the issues of data validity, confidentiality, and cost as they relate to the operations of a data management system in clinical practice.

0094-3509/80/061061-07$01.75

cal and ethical, which arise in the design and implementation of a reliable data collection system in a residency-based ambulatory care setting.

The paper focuses on a computerized data collection system used by the Department of Family and Community Medicine at the University of Massachusetts Medical Center. A program was implemented in April 1977 to document and monitor clinical experiences of family physicians and family practice residents at four affiliated, community based health centers. It is based on an on-line computerized data collection system, by which data pertaining to clinical interactions at each health center are entered into a central computer at the medical center. Data entry is facilitated by the use of checklist encounter forms which are precoded and completed by health care providers. The data are entered by research secretaries at each health center by means of computer terminals. Data items recorded include a patient's chart number, birthdate and sex, date and site of visit, provider code, diagnoses, medications, and referrals.

The data collection system is designed to satisfy three main objectives. The first objective, which is *research* oriented, is to conduct research in the field of family medicine with an emphasis on epidemiology, quantitative analyses of patients and health problems encountered, and patient management. The second objective is oriented towards *education,* and the evaluation of the department's residency training program. The third objective focuses on *service provision,* that is, to supply the health centers with information about their own practices, to aid in administration, practice management, and the research efforts of individual providers.

Factors Impinging on the Validity and Reliability of a Computerized Data Collection System

Central to findings of epidemiological research, based on a large-scale computerized data collection system, are issues of data reliability and validity. In the context of this article, reliability (or reproducibility) is based on the degree to which consistent clinical assessments are recorded when the same clinical circumstances are encountered more than once. The major factor which affects the consistency of results is observer or recorder variation (ie, error). Common examples of such error include: a physician's diagnosing the same condition differently on separate occasions; a physician's recording all diagnoses made in an encounter on some occasions and only the primary diagnosis on other occasions; and keypunching and coding errors made by research secretaries. Validity refers to the degree to which recorded clinical information reflects accurately the clinical circumstances encountered.

As a prerequisite to conducting epidemiologic research, factors which influence data reliability and validity were evaluated six months into the implementation of the system. Factors identified as important include the use of a uniform mechanism for recording clinical data, uniform diagnostic criteria, the identification of sources of recording error, and the attitudes of health care providers to a computerized data base.

Choosing a Mechanism for Coding Data

The choice of a mechanism by which data pertaining to clinical interactions are coded and computerized has a significant impact on the reliability and validity of the information which is collected. Two major options for such coding, with widespread application currently, are diagnostic checklist encounter forms and worksheets/logbooks (eg, E-books in which a provider writes clinical information). At the University of Massachusetts the authors have been able to compare and evaluate both types of recording devices, since one of the four health centers has converted recently from the use of log books to the use of encounter forms. With the change it was found that the average number of diagnoses that were recorded for each patient on a single visit increased significantly from 1.18 to 1.40 diagnoses per visit ($P<.01$). This increased recording could be explained by the shorter time required for a provider to use a diagnostic checklist. It is likely that the increased reporting of diagnoses represents an improvement in data validity because of the recording of more complete diagnostic information.

The use of checklists also improves validity by reducing a major source of error, ie, coding error. (Write-in diagnoses can be miscoded easily by research secretaries, often because it is difficult to find appropriate codes.) However, checklists do introduce biases into the recording process. It is possible that providers could be predisposed to check a precoded diagnosis rather than write in a more precise, but closely related, diagnosis. This problem is ameliorated, somewhat, by having a place on the form for write-in diagnoses. On the basis of the above findings it appears that the incorporation of a checklist encounter form into the data collection process will improve the validity of the data and hence the results of any research based on such data.

Standardization of Recording Mechanisms

Since encounter forms are used as the basis of recording data, it is important to standardize the checklist format for each health center. In this way, the inherent biases that exist when a checklist is used are at least uniform for all health centers and data comparability is ensured. Prior to 1978 each health center was using a different encounter form. This situation arose because of their position as independent clinics prior to their integration into the university's residency program, and their use of different billing companies. Differences in distributions of diseases at each health center could not be attributed solely to variations in the characteristics of either the patient population or treatment practices but possibly, in part, to differences in the design of the encounter forms.* The incorporation of a standardized checklist of diagnostic items into the data collection process at each health center now permits valid comparisons of utilization, diagnostic, and treatment patterns. The determination of diagnoses to be included on the checklist was made by examining frequency patterns of diagnoses made at the four health centers. The pre-coded diagnoses on the encounter form account for 80 to 90 percent of all diagnoses seen at each health center.

*This same limitation applies to the validity of compositive analyses with clinical data recorded in other parts of the United States which are based invariably on unique recording mechanisms.

Uniform Diagnostic Criteria

Another measure which is necessary for recording valid data is the establishment of uniform diagnostic criteria. The experience at the University of Massachusetts has shown providers to have particular difficulty in defining certain organic manifestations of illness, eg, hypertension and its various counterparts (including essential, labile, non-specific, uncomplicated, complicated, elevated blood pressure), and the range of psychosocial and behavioral problems. For example, with respect to behavioral problems, tobacco abuse was diagnosed by some providers if a patient smoked one cigarette a day and by others if a patient smoked one pack a day. The outcome was that these diagnoses were recorded less frequently at these health centers in comparison to national norms. Therefore, a series of seminars involving faculty physicians and residents from all centers is in the planning stages, for the purpose of developing criteria for diagnoses that have divergent interpretations. These activities should provide some local uniformity to the diagnostic process. However, one must realize that comparisons with data collected in other systems may be somewhat invalid due to differences in diagnostic definitions.

Identification of Error

In order to maximize the reliability of data input, it is essential to identify the sources of recording error that exist in the data collection process so that steps can be taken to reduce error in the system. A review of the few documented accounts of error rates in data systems shows that error is measured either in a limited or unclearly delineated manner. Hollison et al,[7] for example, refer only to the rate at which physicians coded data about patients' visits. Marsland et al[1] noted "a four percent recording error between the patient's record problem list, the daily worksheet, and information stored in the computer." The use of

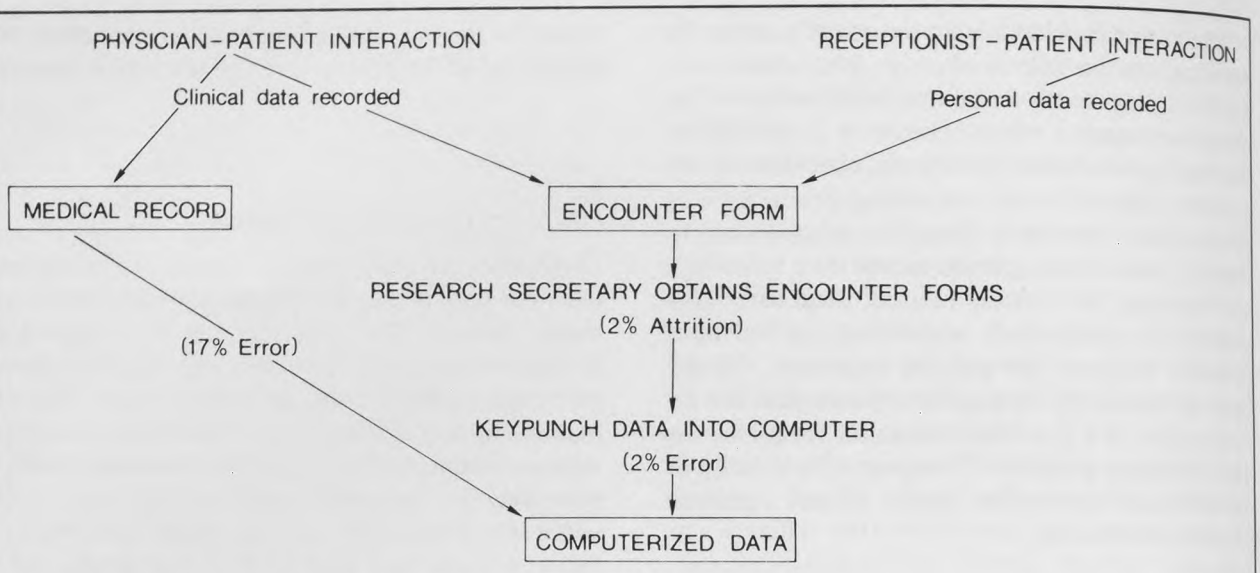

Figure 1. Introduction of error into a computerized data system

three mechanisms for recording means that there are six potential sources of error. There may be an under-reporting or over-reporting of data in the worksheet in comparison to the problem list, an under-reporting or over-reporting of data in the computer in comparison with the worksheet, or an under-reporting or over-reporting of data in the computer in comparison with the problem list. The use of one statistic of error which has six potential sources leaves the reader confused about the exact nature of the error (ie, coding errors or errors of omission) and its source.

With respect to the validity and reliability of data in the present computer system, samples of data were examined to determine:

a. the degree to which patient visits were recorded on encounter forms and entered into the computer

b. the degree to which diagnoses recorded in the computer existed in the medical records and vice versa

c. the degree of keypunching errors in the transference of data from the encounter form to the computer

Figure 1 illustrates the flow of data and the sources and rates of error. At the time of a physician-patient interaction, clinical data, including diagnostic and management information, are recorded by a physician in the medical record and on an encounter form. After this information is recorded by the physician and personal data (patient's name, birthdate, date of visit, chart number) are recorded by a receptionist or nurse, the encounter form is passed on to a research secretary, located at each health center, who enters encounter form data into the computer. An audit of billing records at each health center indicated that 98 percent of all patient visits had encounter forms completed and entered into the computer. The remaining two percent had insufficient data recorded on the encounter form and hence were not entered into the computer as the form was not completed in the first place by the physician. A two-percent error rate was found, also, in keypunching data from the encounter form into the computer.

A final measure of reliability was based on comparisons of diagnoses recorded in medical

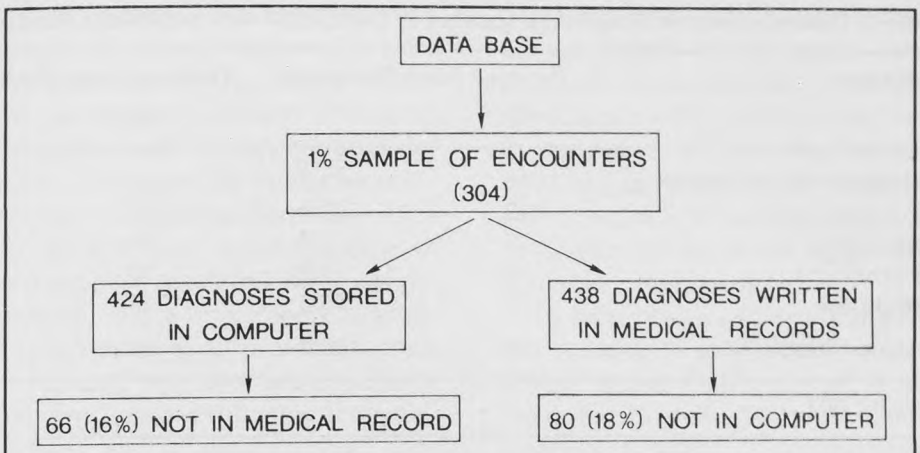

Figure 2. Comparison of diagnoses stored in computer and written in patients' medical records

records with those in the computer. To determine this more detailed assessment of reliability, a one-percent sample (n=304) of encounters made at the health centers during a one-year period was selected randomly. As Figure 1 indicates, on average, there was a 17-percent error rate when comparing diagnoses listed in the medical record with those in the computer. A further breakdown of these data (Figure 2) shows that for 304 patient encounters selected from the health centers, there were 424 diagnoses stored in the computer and 438 entered into the patients' medical records. Of all diagnoses in the medical records, 18 percent were not present in the computer, apparently having been omitted from the encounter form by the physician. Surprisingly, however, of all diagnoses present in the computer, 16 percent were omitted from the medical records. This finding is in contrast to a report of encounter form validity by DeSimone et al,[9] in which the authors state "there were no disease codes recorded on the computer tape that did not correspond to problems noted in the chart."

While discrepancies of the above magnitude should be of concern to researchers, the error rates indicated above are consistent with rates reported elsewhere. For example, Dickie et al[6] noted a 15-percent discrepancy between diagnoses recorded on an encounter form and those recorded on the medical chart. DeSimone et al,[9] on the other hand, reported that only "58 percent of codeable problems listed on the chart actually reached the computer tape."

Furthermore, examination of the types of health problems omitted from the computer and those omitted from the medical records revealed some definite trends in the recording practices of physicians. Diagnoses that were omitted from either source were identified and classified into five broad categories: health maintenance and family planning procedures; acute illness; chronic illness; psychosocial problems; and signs and symptoms (Table 1). Of the 80 diagnoses omitted from the computer (ie, written only in the medical records), 47 percent were signs and symptoms (ICHPPC category XVI). Conversely, of the 60 diagnoses omitted from the medical records (ie, appearing only in the computer and hence on the encounter form), only 20 percent involved signs and symptoms, and the majority were chronic illnesses (42 percent). A chi-square test showed the distributions of the types of diagnoses omitted from the medical records and the types omitted from the computer were significantly different at the .001 level. Having identified the types of health problems that account for the recording errors in these

Table 1. Discrepancies in Diagnoses Located in Computer and in Medical Records

Diagnosis Category	Omitted from Computer		Omitted from Medical Records	
	No.	%	No.	%
Health maintenance, family planning	16	20.3	0	0.0
Acute illness	18	22.8	15	22.7
Chronic illness	6	7.6	28	42.4
Psychosocial problems	13	16.5	10	15.2
Signs and symptoms	37	46.8	13	19.7
Total diagnoses omitted	80	100.0	66	100.0

χ^2 significance: $P<.001$

practices, one can begin to explore with physicians the reasons for their particular recording patterns. If it is possible to determine explanations for the omission of clinical information in the medical record or on the encounter forms, steps can be taken to modify recording patterns.

Gaining Acceptance of the Data System

Another potential obstacle to the recording of complete and reliable data is the degree of acceptance of the system and cooperation by health center providers and administrators. Health care providers' initial reactions towards the implementation of a computerized data collection system may be unfavorable because of required changes in recording habits, and a perceived philosophical inconsistency, among some family physicians, between a less technologically oriented system of health care and the implementation of a computerized data collection system.

A number of measures have been adopted which address such issues. First, instituting a diagnosis checklist, developed in conjunction with the providers themselves, has encouraged the recording of more complete diagnostic information, because of its less time-consuming format. Consequently, it has reduced, somewhat, the particular problem of providers' recording simply the primary diagnosis and neglecting the remainder of health problems discussed during a patient encounter. It is useful, also, to demonstrate the practical uses of the data which providers record in their daily practices. For example, one of the physicians who wanted to conduct an influenza vaccination program in his practice, wished to identify patients that had been diagnosed with chronic heart and lung diseases, so that he could immunize persons at highest risk for the illness. Another physician, interested in issues of continuity of care in his health center, was provided with an appropriate descriptive summary and as a consequence initiated a team concept into his practice, whereby a physician, nurse practitioner, and nurse were assigned to patients as primary care providers. Finally, a Family Oriented Medical Profile is being introduced into each patient's medical chart. The profile provides a computerized summary of the dates of visits, physicians seen, and diagnoses made for all members of a household who made visits to a health center in a two-year period. Thus, it is possible for a physician quickly to assess possible family patterns and dynamics related to illness and health care utilization, and the degree of continuity of family care with respect to the numbers of providers en-

countered. Regular feedback of meaningful data to the providers is an important component in reinforcing the data's applicability. At the University of Massachusetts, descriptive data and written reports are distributed to each health care provider at the health centers on a quarterly basis, for practice management and educational purposes. Each person receives his/her own ranked listing of health problems diagnosed, age-sex profiles of patients and encounters, and a brief report describing the clinical experiences of all providers, comparing health centers. These data are especially useful to medical directors in evaluating the training of their residents, and to administrators concerned with utilization patterns at their center.

Discussion

During the first two years of the data collection process, the authors dealt with many issues in an effort to achieve data reliability and validity, learning to resolve the controversies which occurred occasionally between researchers at the medical school and health center providers. The most significant of these experiences, however, has been the growing realization that there are limitations imposed upon any research effort by factors which are largely out of one's control. These factors can greatly reduce the nature of the data collected and research questions to which it can be applied. Factors of these kinds center on issues such as the personnel and hardware costs involved in establishing a computerized data collection system; confidentiality, which in the University of Massachusetts program required the development of protocols to restrict accessibility to patient information and to safeguard the identity of patients and health centers; computer and programming related issues, which necessitate the advanced planning of data analyses so that data files can be structured accordingly to minimize subsequent programming difficulties.

The factor of most concern, however, and one which has the greatest effect upon the data collection process, ultimately, is the set of health center priorities, which are focused largely on patient care, not research. Because of the semiautonomous nature of the health centers, there are practical restrictions on a researcher's efforts to develop a greater emphasis on research activities. In order to be effective, the data collection process must be a collaborative effort between researchers and providers, an alliance which can be mutually profitable in the areas of patient management, education, and research.

In fact, health center attitudes are changing in this respect, as individual providers and administrators at the University of Massachusetts affiliated health centers are becoming increasingly aware of the utility of the data base and concerned with the quality of the data and its applications. This is in part a result of more regular communication between the medical school and health centers, the dissemination of meaningful and understandable data reports, and the demonstration of the clinical applications of the data base.

References

1. Marsland DW, Wood M, Mayo F: Content of family practice. J Fam Pract 3:23, 1976
2. Newell JP, Bass MJ, Dickie GL: An information system for family practice: Part 1: Defining the practice population. J Fam Pract 3:517, 1976
3. Newell JP, Dickie GL, Bass MJ: An information system for family practice: Part 3: Gathering encounter data. J Fam Pract 3:633, 1976
4. Stewart LC, Gehringer GR, Byars VG Jr: Patient problems in the office practice of six family physicians in Louisiana. J Fam Pract 5:103, 1977
5. Bass MJ, Newell JP, Dickie GL: An information system for family practice: Part 2: The value of defining a practice population. J Fam Pract 3:525, 1976
6. Dickie GL, Newell JP, Bass MJ: An information system for family practice: Part 4: Encounter data and their uses. J Fam Pract 3:639, 1976
7. Hollison RV Jr, Vasquez AM, Warner DH: A medical information system for ambulatory care, research, and curriculum in an Army family practice residency: 51,113 patient problems. J Fam Pract 7:787, 1978
8. Levinson D: Information management in clinical practice. J Fam Pract 7:799, 1978
9. DeSimone JP, Hudson N, Konen JC: Does encounter data really reflect what is seen in the office? J Fam Pract 3:666, 1976