

C. Safran

Center for Clinical Computing,
Beth Israel Hospital,
Harvard Medical School,
Boston, MA, USA

Review Paper

Electronic Patient Records and Clinical Research

Abstract: The digitalization of patient records promises clinicians, researchers, health administrators, and politicians new tools to transform data into information, information into knowledge, and perhaps even knowledge into wisdom. This paper will review the status of clinical databases that can be derived from electronic patient records, discuss how databases derived from routine clinical care are being used, and outline the pitfalls and limitations of using electronic patient records for clinical research.

Keywords: Computer-based Patient Records, Clinical Data Analysis

Introduction

Paper-based medical records are difficult to use for clinical research. They are frequently unavailable because they are being used for patient care, or financial audit, or because they have been misplaced. Even when the medical record can be located, the information in it is poorly organized [54], difficult to read, and time-consuming to abstract. Recognizing these inherent limitations, clinical researchers have usually argued for concurrent, prospective data collection [24,27]. This data collection effort is not meant to replace the clinical record, but rather to produce a research-quality database. This type of approach, used in the rheumatology database ARAMIS [14,15] and the Duke cardiology database [36], is highly successful; it improves our understanding of health care, but requires a high degree of coordination, commitment, and costs.

Clinical computing systems have evolved over the past two decades to the extent that much of the clinical

encounter can be captured in electronic format [1,3,5-7,10,17,29,30,33-35,38,46-48,51,53,55-58]. While researchers can legitimately claim that clinical databases that result from routine collection of data will never be of the same quality as a research database such as ARAMIS or Duke cardiology database [11,12,16,22,25-27,28,44,49,59], many clinicians, health services researchers, administrators, and politicians are beginning to ask questions of these clinical data repositories [2,8,9,13,18,23,26,31,32,40,49,52,59].

To understand the use of routinely collected clinical data for research, one must first distinguish between ambulatory data and acute care data. While the distinction may seem arbitrary from a patient-centered perspective, the mechanisms and incentives for data collection are quite different in the two settings. Moreover, data obtained in the acute care setting are more plentiful and are accumulated over a shorter period. The density of data in the acute care setting makes data analysis easier because there are

fewer missing data items and because the temporal relationships between them may be less important (see section on limitations and bias).

Almost all hospitals worldwide now collect some of their clinical data in electronic form. Traditionally, electronically capture of health data in the United States was stimulated by the need to capture charge data. As the mechanism of reimbursement for acute medical care shifted from transfer of charges to payment linked to diagnosis and procedure codes, so-called claims databases became widely available [22]. With the advent of automated results reporting in the clinical laboratory, clinical information also became more available. Now, with the appearance of less expensive departmental computing, medications and diagnostic results can be stored in electronic form. In the ambulatory care setting, computer-based patient records, first introduced in the 1970's [3,17,29], are now more widely used. In Great Britain [55-57] and the Netherlands [50] government support has led to primary capture of clinical data

in physicians' offices.

These evolving computer-based patient records will solve the problems of access, readability, and perhaps, organization of clinical data [54]. However, computing systems designed to support patient care directly still present problems to those interested in clinical research. First, considerable computer time is needed to aggregate information across patient records [39,41]. Second, and more important, the researcher usually cannot gain access to the clinical database directly; it is necessary to use computer programmers to act as data intermediaries. This situation is analogous to the situation that existed nearly two decades ago, before end-user literature searching was initiated by Bleich [4,20,21]. Moreover, routinely collected clinical data introduce a variety of biases that must be understood before the richness of these clinical databases can be realized [11,12,27,28,44].

Clinical computing systems have traditionally been designed to store and display information about particular patients who are considered active with respect to their care. As a result, data are usually archived once a patient is considered inactive (sometimes in as short a period as 90 days and usually within one year of the patient's last activity). In addition, the data structures necessary to support rapid display of a single patient's data are not optimal for evaluating data in the aggregate. Many hours of computer processing time and much assistance from the programming staff are typically required to obtain information such as the status of all patients with specific laboratory values or x-ray interpretations. Yet the ability to review data from many records can be extremely valuable for the physician or scientist with general or specific questions about patient care. Searches of clinical data could also help the clinician decide on a course of action for a particular patient with unusual clinical findings

[23]; patients with clinical similarities can be identified and their medical records examined for precedence for success in therapy [8]. Whether these data are stored in the databases of departmental computers or in a central data repository, they represent a potential gold mine of information and knowledge. Rapid access to such information can help bridge the gap between clinical care and research, support clinical and executive decision making, and improve the quality of care [19,28,32,37,41,44,49,58].

Uses of Clinical Data

A clinical database can be used in four ways: (1) to display information about an individual patient (results reporting), (2) to find data on a patient with similarities to one being seen (case finding), (3) to describe a group of patients with at least one attribute in common (cohort description), and (4) to analyze data patterns for trends or relationships (predictive modeling).

The first use of clinical data is the most common, and is at the heart of every clinical information system. Rapid access to data on individual patients is the primary informational need of every practicing clinician. At Boston's Beth Israel Hospital, clinicians look up individual patient information in the computing system 100 times more frequently than they look up data in aggregate [39].

In the second instance, a clinician might wonder, "Have we ever seen a patient like this before? I remember seeing a patient two years ago with AIDS and pericarditis who..." or "Could I find a recent case for teaching rounds of a patient with diabetic ketoacidosis with an anion gap greater than 24 who...?" While case finding is obviously useful, problems relating to pattern matching, often with incomplete data, are formidable. Moreover, the more precise the specification, the

larger the clinical database needs to be to contain probable matches [8].

In the nine years that clinicians at Boston Beth Israel Hospital have been searching the ClinQuery database [39], cohort description has been the most frequent reason. In this instance, clinicians and health administrators ask, "Find me all the patients with a diagnosis of AIDS and describe this population in terms of their demographics and resource utilization." Cohort description (data aggregation) is the first step in transforming clinical data into information. Both case finding and cohort description are exploratory in nature and require repetitive interaction with the clinical database [43].

The fourth use of clinical data focuses on prediction. The questions arise, "How can we use the richness of our evolving clinical data repositories to derive relationships or hypotheses?" How can clinical data be transformed into knowledge? Statistical approaches to parting and analyzing clinical records can be used to answer such questions as "Can we identify patients at risk for bad outcomes?" [18,26,31,49], but the general problem of prediction remains difficult. However, this use of clinical data should be our goal.

Limitations and Biases

It seems unlikely that many important clinical questions will be subject to randomized clinical trials because of the ethics, logistics, and expense that would be involved [16,27]. Evolving statistical and epidemiological methods allow us to approach these clinical data repositories with the purpose of building predictive models, but a clear understanding of the limitations of routinely collected clinical data and the inherent biases is necessary [12,28,32,37,44,50,59].

Perhaps the most pressing question about routinely collected clinical data

concerns quality [11,22]. Do these data have the same quality as research databases like ARAMIS or the Duke cardiology database? Data collected for a particular reason, such as charge capture, may be accurate enough for this designated purpose, but inaccurate for other reasons. The diagnosis assigned to an episode of care is a good example of these limitations. Even if the clinician is selecting a diagnosis code and there is sufficient vocabulary to allow an appropriate diagnosis, the clinician may be intentionally misleading. For example, a clinician might assign a diagnosis of "irregular menses" rather than "fertility control," if the diagnosis was needed only for a billing form that would be sent to a young woman's parents.

Organizing and collecting clinical data is costly, and this cost introduces a collection bias. While clinical care and research were the early reasons for organization of data, the fiscal imperatives of the past two decades have governed the development of most health care computing. Thus, data have been captured as if doing so were equivalent to charge capture.

In many instances, the collected data, such as discharge diagnoses, are fraught with inaccuracies [11,22,40]. Both errors of omission and commission occur when hospital personnel assign codes to patient encounters. These codes can be random or systematic, depending on the diagnosis, and error rates are frequently reported to exceed 20%. Having access to clinical data as an alternative to abstracted codes circumvents this problem in some instances.

Not all patients have the same data collected for each episode of care. The recognized wide variability in practice styles, as well as the underlying care issues, leads to a selection bias in data reporting. In other words, sicker patients may have more testing. Thus, analysis of clinical databases must not

only recognize these inherent biases, but also effectively deal with the complexity of missing data.

As access to data sources increases, the ability to link patient records presents new problems for clinical researchers. If patients can be uniquely identified in two databases, are data from one institution comparable to data from another? For the individual test, were they conducted on the same machine, with the same technique, and reported with the same normal values? In reality, every institution that seeks to build a clinical data repository has to consider this issue as its database accumulates over time and its laboratories change and update their procedures. However, when two databases from different institutions are centralized in a common data repository, this issue must be resolved. Even without considering common patients between two institutions, the comparability of data elements must be taken into account. Standardized coding of diagnoses and other patient information is just one of the issues that need to be addressed.

The promise of centralized clinical data repositories is that patient records will not only be linked, but also followed longitudinally. Time-oriented representation of data and time-linked analysis are critical to making good use of clinical data [15]. For instance, suppose a clinician wants to compare the diabetic control of two populations of patients. Since every patient will have a different period of observation, this time-dependent relation will have to be analyzed.

Because all data collected from patients come at a cost, even if it is simply the clinician's time, data may not always be collected by current clinical computing systems. In acute care settings, for example, it may be difficult to know which clinicians are responsible for a patient's care.

Discussion

The computer-based patient record linked to longitudinal clinical data repositories represents a valuable resource for decision makers (clinicians, hospital administrators, and health policy executives). Technological advances and improvement in the cost-performance of random-access storage now make it possible to consider large-scale clinical databases that are national or international in scope. If ethical considerations such as confidentiality of information can be assured [39,42], these repositories should advance both clinical medicine and health policy. The largest barrier to using routinely collected clinical data is not the limitations of the data themselves, but rather the lack of a data paradigm for the decision maker [44]. Decision makers are well trained to recognize when they lack sufficient data; however, they are less well trained to work with data when the data are provided.

Although the use of routinely collected clinical data has limitations, use of aggregated data collected from large numbers of computer-based patient records is virtually limitless. Today's clinical researchers have many of the same problems facing radio-astronomers [28]. Clinical information systems are collecting vast quantities of data. Some of these data may be inaccurate, and some are merely noise. We need to filter out the noise and transform data to information and then to knowledge.

References

1. Bakker AR. The development of an integrated and co-operative hospital information system. *Med Inf* 1984; 9:135-42.
2. Barbash GI, Safran C, Ransil BJ, Pollack MA, Pasternak PC. Need for better severity indexes of acute myocardial infarction under diagnosis-related group. *Am J Cardiol* 1987; 59:1052-8.

3. Barnett GO. Computer-stored ambulatory record (COSTAR). NCHSR Research Digest 1976; DHEW publication HRA 76-3145.
4. Beckley RF, Bleich HL. Paper Chase: a computer-based reprint storage and retrieval system. *Comp Biomed Res* 1977; 10:423-30.
5. Bleich HL, Beckley RF, Horowitz G, et al. Clinical computing in a teaching hospital. *N Engl J Med* 1985; 312:756-64.
6. Bleich HL, Safran C, Slack WV. Departmental and laboratory computing in two hospitals. *MD Comput* 1989; 6:149-55.
7. Campbell JR, Civner N, Seelig CB, et al. Computerized medical records and clinic function. *MD Comput* 1989; 6:282-7.
8. Chute CG. Clinical data retrieval and analysis: I've seen a case like that before. *Ann New York Acad Sc* 1992; 670:133-40.
9. Degoulet P, Chatellier G, Devries C, Lavril M, Menard J. Computer-assisted techniques for evaluation and treatment of hypertensive patients. *Am J Hypertension* 1990; 3:156-63.
10. Degoulet P, Menard J, Berger C, Plouin PF, Devries C, Hirell JC. Hypertension management: the computer as a participant. *Am J Med* 1980; 68:560-7.
11. Demlo LK, Campbell PM, Brown SS. Reliability of information abstracted from patients' records. *Med Care* 1978; 16:995-1005.
12. Feinleib M. Data bases, data banks and data dredging: the agony and the ecstasy. *J Chron Dis* 1984; 37:783-90.
13. Fischer KF, Lees JA, Newman JH. Hypoglycemia in hospitalized patients: causes and outcomes. *N Engl J Med* 1986; 315:1245-50.
14. Fries JF. ARAMIS Second Progress Report. In: *American Rheumatism Association Medical Information System*, Palo Alto, CA 1978.
15. Fries JF. Time-oriented patient records and a computer-databank. *JAMA* 1972; 222:1536.
16. Grémy F. Informatics and medical methodology: random reflections about clinical data bases. *Med Inform* 1982; 7:85-91.
17. Grossman JH. An ambulatory medical record system for patient care and health care management. *Methods Inf Med (Suppl)* 1972; 6:375-82.
18. Herrmann FR, Safran C, Levkoff SE, Minaker KL. Serum albumin on admission as a predictor of death, length of stay, and readmission. *Arch Intern Med* 1992; 152:125-30.
19. Hlatky MA, Lee KL, Harrell FE, et al. Tying clinical research to patient care by use of an observational database. *Stat Med* 1984; 3:375-87.
20. Horowitz GL, Bleich HL. PaperChase: a computer program to search the medical literature. *N Engl J Med* 1981; 305:924-30.
21. Horowitz GL, Jackson JD, Bleich HL. PaperChase: self-service bibliographic retrieval. *JAMA* 1983; 250:2494-99.
22. Iezzoni LI, ed. *Risk Adjustment for Measuring Health Care Outcomes*. Ann Arbor, MI: Health Administration Press, 1994.
23. Klein LA, Compton CC. A 51-year-old woman with microscopic hematuria and an osteolytic lesion of the spine. *N Engl J Med* 1986; 314:500-9.
24. Kurland LT, Molgaard CA. The patient record in epidemiology. *Sci Am* 1981; 245:54-63.
25. Lee TH, Goldman L. Development and analysis of observational data bases. *J Am Coll Cardiol* 1989; 14:44A-7A.
26. Levkoff SE, Safran C, Cleary PD, Gallop J, Phillips RS. Identification of elderly patients at risk for delirium. *J Am Geriatr Soc* 1988; 36:1099-104.
27. Mantel N. Cautions on the use of medical databases. *Stat Med* 1983; 2:355-62.
28. McDonald CJ, Hui SL. The analysis of humongous databases: problems and promises. *Stat Med* 1991; 10:511-20.
29. McDonald CJ, Blevins L, Tierney WM, Martin DK. The Regenstrief medical records. *MD Comput* 1988; 5:34-47.
30. McDonald CJ, Murray R, Jeris D, et al. A computer-based record and clinical monitoring system for ambulatory care. *Am J Public Health* 1977; 67:240-5.
31. Phillips RS, Safran C, Cleary PD, Delbanco TL. Predicting emergency readmissions for patients discharged from the medical service of a teaching hospital. *J Gen Intern Med* 1987; 2:400-5.
32. Pryor DB, Califf RM, Harrell FE, et al. Clinical data bases: accomplishments and unrealized potential. *Med Care*. 1985; 23:623-47.
33. Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. *J Med Syst* 1983; 7:87-102.
34. Pryor TA. The HELP medical record system. *MD Comput* 1989; 5:22-33.
35. Rector AL, Nowlan WA, Kay S. Foundation for an electronic medical record. *Methods Inf Med* 1991; 30:179-86.
36. Rosati RA, McNeer JF, Starmer CF, Mittler BS, Morris JJ Jr, Wallace AG. A new information system for medical practice. *Arch Intern Med* 1975; 135:1017-24.
37. Rosati RA. Tying the output of clinical research to patient care. *Proc Annu Meet Med Sect Am Counc Life Insur* 1976; 1:47-60.
38. Safran C, Bleich HL, Slack WV. Role of computing in patient care in two hospitals. *MD Comput* 1989; 6:141-8.
39. Safran C, Porter D, Lightfoot J, et al. ClinQuery: a system for online searching of data in a teaching hospital. *Ann Intern Med* 1989; 111:751-6.
40. Safran C, Porter D, Slack WV, Bleich HL. Diagnosis-related groups: a critical assessment of the provision for comorbidity. *Med Care* 1987; 25:1011-4.
41. Safran C, Porter D. New uses of the large clinical data base at the Beth Israel Hospital: on-line searching by clinicians. In: *Proceedings of the 10th Annual Symposium on Computer Applications in Medical Care*. Washington DC, 1986:114-9.
42. Safran C, Rind D, Bakker AR, Citroen M, Slack V, Bleich HL. Protection of confidentiality in the computer-based patient record. *MD Comput* 1995 (in press).
43. Safran C, Sobel E, Lightfoot J, Porter D. A computer program for interactive searches of a medical data base. In: Salamon R, Blum B, Jorgenson M, eds. *MEDINFO 86: Proceedings of the fifth world conference on medical informatics*. Amsterdam: Elsevier Science Publishers, 1986:545-9.
44. Safran C. Using routinely collected data for clinical research. *Stat Med* 1991; 10:559-64.
45. Sands DZ, Safran C, Slack WV, Bleich HL. Use of electronic mail in a teaching hospital. *Proc Annu Symp Comput Appl Med Care* 1993; 306-10.
46. Scherrer JR, Baud R, Brisebarre, Messmer E, Assimacopoulos A. A hospital information system in continuous operation and expansion: concepts, tools and migration. In: Othner HF, ed. *Proceedings of the 10th Annual Symposium on Computer Applications in Medical Care*. Washington DC 1986:120-32.
47. Scherrer JR, Baud RH, Hochstrasser D, Ratib R. DIOGENE: an integrated hospital information system. *MD Comput* 1990; 7:81-9.
48. Stead WW, Hammond WE. Computerized medical records: a new resource for clinical decision making. *J Med Syst* 1983; 7:213-20.
49. Tierney WM, Martin DK, Greenlee MC, et al. The prognosis of hyponatremia at hospital admission. *J Gen Intern Med* 1986; 1:380-5.

50. Tierney WM, McDonald CJ. Practice databases and their uses in clinical research. *Stat Med* 1991; 10:541-59.
51. Van der Lei J, Duisterhout JS, Westerhof HP, et al. The introduction of computer-based patient records in the Netherlands. *Ann Intern Med* 1993; 119:1036-41.
52. Van der Lei J, Musen MA, van der Does E, Man in't Veld AJ, van Bommel JH. Comparison of computer-aided and human review of general practitioners' management of hypertension. *Lancet* 1991; 338:1504-8.
53. Warner HR, Olmsted CM, Rutherford BD. HELP: a program for medical decision-making. *Comput Biomed Res* 1972; 5:65-74.
54. Weed LL. Medical records that guide and teach. *N Engl J Med* 1968; 278:593-600.
55. Whiting-O'Keefe QE, Simborg DW, Epstein WV, Warger AW. A computerized summary medical record system can provide more information than the standard medical record. *JAMA* 1985; 254:1185-92.
56. Wyatt J. Clinical data systems, part 1: data and medical records. *Lancet* 1993; 344:1543-7.
57. Wyatt J. Clinical data systems, part 2: components and techniques. *Lancet* 1993; 344:1609-14.
58. Wyatt J. Clinical data systems, part 3: data and medical records. *Lancet* 1993; 344:1682-8.
59. Zwetsloot-Schonk JHM, Snitker P, Vandenbroucke JP, Bakker AR. Using hospital information systems for clinical epidemiological research. *Med Inform* 1989; 14:53-62.

Address of the author:
 C. Safran
 Center for Clinical Computing,
 Beth Israel Hospital,
 Harvard Medical School,
 350 Longwood Avenue,
 Boston, MA, 02115,
 USA