Obstacles and Approaches to Clinical Database Research: Experience at the University of California, San Francisco

Thomas B. Newman, MD,MPH
Andrew Brown MD, MPH
M. Janet Easterling

Department of Laboratory Medicine, University of California, San Francisco, CA

With increasing availability of clinical data in machine-readable form, and decreasing cost of storing and manipulating that data, retrospective research using clinical databases has become more feasible. Nonetheless, much of the potential for clinical research using these data remains unrealized. Obstacles to clinical database research include difficulty accessing data, difficulty using retrospective data to draw valid inferences about medical tests and treatments, and a shortage of investigators trained and interested in using a clinical database to answer their questions. At the University of California, San Francisco, we have developed a Clinical Database Research Program (CDRP) to try to overcome these obstacles. The CDRP maintains a relational database of patient data obtained from diverse sources and a small staff dedicated to providing such data to researchers. The CDRP staff also provides support for design and analysis of studies using the database--the development of methods for such studies is our primary research interest. Finally, to increase the number of investigators using the database for research, we are integrating training in clinical epidemiology and clinical research methods into residency and fellowship training, and offering an elective in clinical database research for trainees who wish to undertake a specific project.

As more and more patient care activities generate data in machine readable form, and available hardware and software improve, there is increasing potential for using routinely collected clinical data for research [1-3]. However, considering the vast stores of clinical data theoretically available to potential investigators, the actual amount of clinical research therefrom has been quite modest. In this paper, we will review some obstacles to clinical database research and the approaches the University of California, San Francisco (UCSF) Clinical Database Research Program (CDRP) is taking to overcome them.

OBSTACLES

Difficulty Accessing Data

Data are dispersed. At UCSF, there is a wide variety of clinical data already in machine-readable form. However, as at many medical centers [4], the data are dispersed in separate systems that have evolved independently of one another. Independent computer systems have been implemented in the clinical laboratory, medical records, the pharmacy, radiology, the emergency department, and so on. Some of these systems provide data to the medical center's clinical display system (STOR), which allows retrieval of data on a particular patient at the point of care. However, what is sent to the clinical display system is often a free text report. Coding and data structure present in the original database are not generally preserved. In addition, many data are currently not reported to STOR. Thus, at UCSF useful data are difficult to obtain because of the number of different platforms and personnel involved.

Competition between research and patient care. That relevant data are scattered throughout the medical center is not necessarily an insurmountable obstacle to clinical researchers. For many research questions, data from only one or a few sources might be enough. However, simply finding the data is not sufficient--the personnel familiar with the systems involved have to agree to provide it (in suitable format, with appropriate documentation) to potential investigators. These personnel often do not view provision of data for research as their highest priority. Furthermore, clinical database research is often an iterative process. Thus, the first (and second and third) request for data may need to be revised after some initial analyses. This process is difficult if the personnel involved do not view provision of research data as an important part of their job. In addition, the competition is not just

0195-4210/94/$5.00

for personnel, but for computer time. In order to avoid slowing down clinical systems, research queries may need to be run as batch jobs late at night, diminishing the possibility of interactive data requests.

Scientific Problems with Retrospective Studies

Data collected as a part of patient care are less suitable for answering many types of research questions than data collected prospectively, either for an observational study or for part of a clinical trial [5]. For example, whereas most guidelines for designing or interpreting studies of diagnostic tests specify that a "gold standard" test must be uniformly and blindly applied to all the subjects, in clinical practice results on one test often affect the decision to order subsequent tests, as well as their interpretation. Similarly, retrospective studies of treatments are difficult because the intensity of treatment for a particular disease is likely to be correlated with a worse outcome from that disease simply because patients with more severe disease are treated more intensively. These scientific problems with retrospective studies limit the types of questions that can be addressed with a clinical database, and tend to be discouraging to potential investigators, particularly those without advanced training in epidemiology or statistics.

Underutilization of data

Research is time consuming. Even when the data collection phase is abbreviated by accessing existing data, considerable effort is required to review the literature, design a study, analyze the data, write the paper and get it published. Thus, if only investigators closely associated with a clinical database are mining it for research purposes, it is likely to be underutilized. To realize the full potential of clinical databases for research, the number of investigators using them should be maximized.

There are, however, a number of obstacles to attracting investigators to this type of research. In academic medical centers, tradition, funding availability, and prestige all tend to focus investigators' efforts on becoming an expert on a narrow topic. For any particular topic, the amount that one can learn from a clinical database is limited. The promise of clinical database research is a small (but significant) amount of information about a multitude of different topics, rather than the great depth of study about a particular topic that is helpful for long-term funding and academic success. Thus one of the problems in attracting investigators to clinical database research is that faculty see it as unlikely to advance their careers.

APPROACHES

Facilitating Access: the CDRP Database

An obvious solution to dispersed data is integration. Other investigators have reported systems for integration of heterogenous databases that involve querying the component databases nightly [4], or at the time of a query [6]. At UCSF we took a different approach, dictated by different goals and limited resources. We wanted to allow interactive access to the database while avoiding competition for personnel or computing time between research and clinical care. In addition, because the primary purpose of our database is retrospective research, efficient access to a large store of historical data was more important to us than immediate access to data that are current or even several months old. We therefore obtained data, usually in the form of formatted ASCII text, from a variety of computers on campus, and placed the data into a separate relational database that can be queried interactively for research. We update each of the various components of the database about twice a year from dumps from the computers on which the current data reside.

An overview of the most important current contents of the database, together with approximate space requirements, is provided in the Table. Most of the data are numerical or coded, but some laboratory results are free text. To facilitate retrieval of laboratory data, we used a fourth generation language (4GL) to add unique admission numbers and binary flag fields for maximum, minimum, first, and last for each inpatient test result to identify those results most likely to be of interest from a particular admission.

In spite of not doing any primary data collection ourselves, creation of the database has been labor intensive. The current database, which includes about 5.5 years of data, has taken about two person-years to assemble. Most of this has been programming time, but identifying data sources and getting the personnel involved to provide the data dumps is also time-consuming. In

Table: Major tables in the UCSF-CDRP database, as of August, 1994.

Source of Data Table	# Records	Storage Space	Example fields
Clinical laboratory			
Results	31,000,000	3,200 Mb	Unit number, date, test number, result text, numeric result, admit number, maxflag (see text)
Microbiology			
Specimens	549,000	250 Mb	Unit number, specimen number, source, date
Isolates	163,000	13 Mb	Specimen number, organism, count
Medical records			
Patients	239,000	37 Mb	Name, unit number, sex, date of birth
Admissions	175,000	30 Mb	Admit number, admit date, drg
Diagnoses	688,000	52 Mb	Admit number, ICD-9 diagnosis code
Procedures	388,000	35 Mb	Admit number, ICD-9 procedure code
Obstetrics	29,000	6 Mb	Maternal age, delivery date, type of delivery, gestational age, birth weight
Cardiology			
Echocardiograms	280,000	17 Mb	Unit number, procedure number,
Treadmill tests	37,000	8 Mb	procedure date, result code, admit number

addition, to assure the completeness and quality of the data we have performed regular comparisons with written medical records. However, now that this considerable amount of ground work has been accomplished, we will be able to turn increasing attention to adding data from new sources, to facilitating retrieval with graphical interfaces, and to recruiting and assisting investigators (see below).

Rigorous Retrospective Studies

Although there are many questions that cannot be answered using retrospective data, there are many that can. Because our database is particularly rich in diagnostic data, we have focussed on study designs for assessing diagnostic tests.

Most guidelines on design or interpretation of studies of diagnostic tests are aimed at protecting against falsely concluding that a diagnostic test is helpful when it is not. This is because traditionally, many more research studies (and papers) have been directed at identification of new diagnostic tests than at evaluation of existing tests. The times, however, are a changin'. In the era of managed care, there is increasing interest in identifying existing diagnostic tests that are *not* useful. Luckily, it is easier to show that a test is *not* useful than that it *is* useful. For a diagnostic test to be clinically useful, it must be abnormal at least some of the time, these abnormalities must not be readily predictable from other available data, abnormalities must affect management, and the management decisions must lead to a better outcome. A study that calls into question any one of these necessary but not sufficient criteria can suggest that a diagnostic test is not useful.

One study design that lends itself well to clinical database research is what we have called a *Diagnostic Yield Study*. This design is appropriate for examining tests for diseases that are often sought and seldom found. (Such tests seem to be done frequently at academic centers.) In a diagnostic yield study, a group of patients of

interest is identified based on their having had a particular laboratory test. In many cases, this group can be made more homogenous with respect to clinical indications for the test by including or excluding patients based on results of other tests, previous discharge diagnoses, and so on. Results on the test for the whole group of subjects are obtained, but only those with abnormal results are studied further. The questions such a study can answer are: how often is the test abnormal? When it is abnormal, could the abnormality have been predicted from other tests? Was management affected by the abnormal result? What was the outcome?

For example, we used the database to study the diagnostic yield of direct bilirubin levels in jaundiced newborn babies [7]. From about 5000 determinations of direct bilirubin, we identified those whose results were above the 95th percentile. In most cases we could see from the database that the result was not clinically significant because the elevation was temporary and was not accompanied by any other tests or discharge diagnoses suggestive of hepatobiliary disease. In the relatively few cases in which there was any doubt, we reviewed medical records. All of the infants who appeared to have had clinically significant direct hyperbilirubinemia had other signs or laboratory evidence of illness before the direct bilirubin elevation was noted. We recommended that direct bilirubin levels be ordered much more selectively in jaundiced infants.

We have done similar studies (with similar results) on the other laboratory tests commonly done to evaluate jaundice in newborn babies [8]. Other examples of diagnostic yield studies that have identified unnecessary or over-used tests include a study of IgM levels as a screening test for congenital infection in small-for-gestational-age newborns [9], and a University of Pennsylvania study of stool cultures for hospital-acquired diarrhea [10].

A variant of the diagnostic yield study is a study of *diagnostic redundancy.* In this design, results of two or more tests ordered together are compared to identify how often they give discrepant results. (When both are normal or both are abnormal, it presumably was not necessary to do both tests.) Medical records of the small subset with discrepant results can then be reviewed to see which test (apparently) gave the right answer. Even in cases where no "gold standard" is available, records can be reviewed to determine which test was *believed* by the treating physicians, and whether specific circumstances can be defined in which one test or the other is more likely to be helpful.

For example, at UCSF a "liver panel" included determinations of aspartate amino transferase (AST), alkaline phosphatase, and total bilirubin. However, using the database we found that there were very few instances in which the bilirubin was high when both the AST and alkaline phosphatase were normal, and that these instances seldom reflected liver disease [11]. As a result, bilirubin was removed from the liver panel.

Similarly, a Clinical Scholar is using our database to determine how often Lactate Dehydrogenase (LD) isoenzymes provide information beyond that available from creatinine kinase (CK) isoenzymes in the diagnosis of acute myocardial infarction. By selecting patients in whom CK and LD isoenzymes are ordered on the day of admission, patients in whom the goal is to rule-out (or in) a myocardial infarction (MI) can be readily identified. In most such patients, results of both CK and LD isoenzymes are congruent--either both normal or both abnormal. In the minority in whom there is a discrepancy, discharge diagnoses can be reviewed, to determine which result was believed. (Although our database includes discharge diagnoses, our experience is that they are not abstracted reliably enough to use as an outcome variable, so they must be obtained from the medical record for a study of this design.) If circumstances in which LD is believed over CK can be identified (e.g., when the duration of symptoms before admission is longer or when, based on other studies, the CK seems to be falsely negative), we can generate recommendations for more selective ordering of that test.

Recruiting and Training Investigators

A major challenge for those interested in use of clinical databases for research is attracting investigators to use the data. Creation and maintenance of a clinical database are expensive. To justify this investment, the number of questions the database is used to answer must be maximized. This means, for the most part, maximizing the number of people using it.

The first step in encouraging investigators to use a clinical database for research is to assemble a database that includes data items of sufficient interest. At present, the richest, most reliable data in the UCSF CDRP are data from the clinical labs. Our strategy is to obtain additional data according to the interests of potential investigators. We obtained cardiology and obstetrics data because we had fellows interested in using them. Our next two additions to the database are likely to be mortality data and pharmacy data. These will greatly expand the range of possible studies that can be done with the database.

As discussed above, a major obstacle for investigators is that many of the studies for which use of a clinical database is most feasible are relatively small clinical studies, not likely to attract extramural funding. On the other hand, ready availability of data for modest studies is very helpful for trainees. Using a clinical database to do one's own research project is a great way to learn about clinical research. Thus one large group of potential users includes medical students, residents, and fellows who are interested in clinical research. The possibility that, if they think of a good question, they can design and begin a study in one month, and possibly finish in another provides strong motivation. We are developing an elective in clinical database research that will help lead trainees through the steps involved - from identification of a research question to writing up the results.

CONCLUSION

Clinical database research has great potential, not just for answering clinical questions, but as a tool for training residents, fellows, and faculty in clinical research. Centralizing patient data in a relational database that allows interactive queries, providing a staff knowledgeable in study design and analysis, and facilitating clinical database research by trainees seem to be promising ways to realize more of this potential.

References

1. Pryor DB, Califf RM, Harrell F Jr., et al. Clinical data bases. Accomplishments and unrealized potential. Med Care 1985;23:623-47

2. Tierney WM, McDonald CJ. Practice databases and their uses in clinical research. Stat Med 1991;10:541-57

3. Safran C. Using routinely collected data for clinical research. Stat Med 1991;10:559-64

4. Marrs KA, Steib SA, Abrams CA, Kahn MG. Unifying heterogeneous distributed clinical data in a relational database. Proc Annu Symp Comput Appl Med Care 1993;644-8

5. Mantel N. Cautions on the use of medical databases. Stat Med 1983;2:355-62

6. Kamel MN, Zviran M. Heterogeneous databases integration in a hospital information systems environment: a bottom-up approach. Proc Annu Symp Comput Appl Med Care 1991;363-7

7. Newman TB, Hope S, Stevenson DK. Direct bilirubin measurements in jaundiced term newborns. A reevaluation. Am J Dis Child 1991;145:1305-9

8. Newman TB, Easterling MJ. Yield of reticulocyte counts and blood smears in term infants. Clinical Pediatrics 1994;33:71-6

9. Mahon BE, Yamada E, Newman TB. Problems with serum IgM as a screening test for congental infection. Clinical Pediatrics 1994;33:142-6

10. Siegel DL, Edelstein PH, Nachamkin I. Inappropriate testing for diarrheal diseases in the hospital. JAMA 1990;263:979-82

11. Brown AN, Sheiner LB, Cohen SN. Evaluation of bilirubin in a liver screening panel [letter]. JAMA 1992;268:1542