Original Articles

Pathology Information Systems

Data Mining Leads to Knowledge Discovery

Jay M. McDonald, MD; Stephen Brossette; Stephen A. Moser, PhD

• Information systems in pathology provide opportunities for pathologists and clinical laboratory scientists to impact both clinical care and modern research agendas. The paradigm shift in health care from individualized care to population-based and standardized delivery systems has created both of these opportunities. In research, pathology information systems can provide key databases for health services research and new informatics-based approaches to database research. The latter is characterized by utilization of pathology databases for data mining to discover new patterns that provide new knowledge. The multidisciplinary knowledge discovery and data mining program at the University of Alabama at Birmingham focuses on this health care application, which has the potential to make a major impact on health care research and delivery.

(*Arch Pathol Lab Med.* 1998;122:409–411)

Pathology information systems represent an area in which the discipline of pathology has an opportunity to add real value to the health care system, both at local and national levels. In addition to adding value in the delivery of quality health care, pathology as a discipline has the opportunity to use information systems for targeted high-impact research. Numerous authors have promoted and documented various aspects of how this can and should occur.[1–3] At the Association of Pathology Chairs meeting in July 1996, a break-out discussion group identified challenges and opportunities for laboratory medicine. Included among the eight major points were many components that focused on information systems and informatics, such as managing clinical research databases, on-line interactive clinical resource utilization management, creation of laboratory-based alerts at order entry, general application of medical informatics, and involvement in telemedicine.

The new health care era, which is dominated by managed care and is focused on cost-effective delivery systems, has led a major paradigm shift in medical care that places patient information systems and databases at the core of the activity. This paradigm shift, from individual patient disease management to population-based disease management, is being accompanied by a shift from high- to low-cost centers and providers (eg, from inpatients to outpatients and from MD providers to health care team providers, often nurse specialists and physician assistants).

The need for making the required changes is obvious. We have entered an era of evidence-based medicine. That is, clinical decisions must be based on scientific evidence that demonstrates effectiveness.[4] Well-trained pathologists and clinical laboratory scientists, as scientifically trained independent evaluators, have both an opportunity and an obligation to paticipate in this process. Outcomes assessment becomes the basis for evaluating appropriateness of clinical decisions and developing population-based clinical guidelines. Feussner[4] has provided an excellent objective approach to evaluating scientific clinical evidence. In addition, he provided an overall spectrum for assessing patient outcomes (Table 1). The pathology information system databases are of enormous value in establishing and evaluating clinical guidelines, the development of which will dominate the medical agenda for many decades. The development and implementation of clinical guidelines is a standard quality improvement process (Figure). Information system databases are often of value in establishing the plan, but are usually of even more value in the measurement/continuous improvement phase.

Providing support via pathology information systems for outcomes management often takes the form of utilization management or "gatekeepers." Either retrospectively, or prospectively, the laboratory information system can be used to assist in enforcing best care practices. The spectrum of activities ranges from preventing repeat orders within defined time frames to monitoring the rate of obtaining interpretable Papanicolaou smears from physicians' offices. These activities are valuable within most health care environments and for managed care and insurance companies.

PATHOLOGY INFORMATION SYSTEMS AND OUTCOMES RESEARCH

The laboratory medicine–specific health services research agenda (which we consider analogous with the out-

Accepted for publication January 20, 1998.

From the Department of Pathology, The University of Alabama at Birmingham (Drs McDonald and Moser), and Computer Information Sciences, The University of Alabama at Birmingham (Mr Brossette).

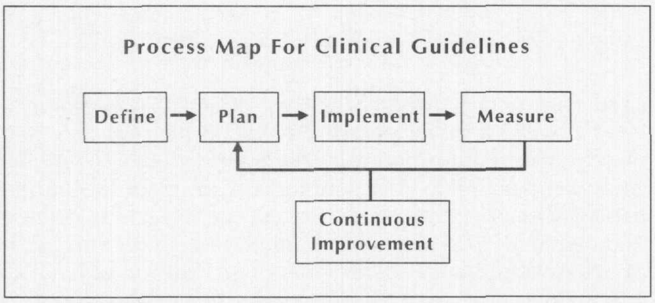
Presented at the College of American Pathlogists Foundation Conference XIII, Patient-Centered Pathology Practice: Outcomes and Accountability, February 1, 1997.

Reprint requests to Department of Pathology, University of Alabama at Birmingham, 519 Lyons-Harrison Research Bldg, 7015 19th St, Birmingham, AL 35233-7331 (Dr McDonald).

Table 1.—Spectrum of Patient Outcomes*

Mortality	
	Crude mortality rate
	Cause-specific rates
Morbidity	
	Comorbid diseases
	Disease severity indices
	Other adverse events, eg, nosocomial infections
Quality of life	
	General or disease-specific
Functional status	
Resource utilization	
Cost of care	

* Adapted from Feussner.[4]



General overview of the clinical guideline development and implementation process.

Table 2.—Important Variables That Affect Outcomes in Laboratory Medicine Health Services Research*

Provider	
	Self
	Paramedical practitioner
	Primary care physician
	Specialist physician
Service	
	Laboratory analyses
	Quality assurance
	Utilization management
	Information system changes
	Analysis of aggregate or individual patient data
Location	
	Home
	Hospice
	Nursing home
	Pharmacy
	Office
	Clinic
	Hospital

* Adapted from McDonald et al.[5]

comes research agenda) has some unique components distinct, at least in part, from the agenda of the overall outcomes research agenda for health care. The pathology agenda specifically focuses on the value of laboratory analyses and pathology services in the delivery of cost-effective, high-quality medical care. One role of pathology information systems in the broader agenda involves integration of large disparate databases. Successful performance of pathology health services research includes three principles: (1) the pathologist must leave the laboratory and function in other administrative and clinical arenas of the health care system, (2) the pathologist must lead and participate in multidisciplinary teams (all outcomes investigations are multidisciplinary); and (3) investigations must be hypothesis-driven. Omission of the last item is a common and very serious error. If a hypothesis is not tested, it is unlikely that conclusions can be extrapolated to future endeavors. There are many key variables that must be assessed when performing laboratory medicine health services research. As shown in Table 2, the impact of the provider, the service provided, the analysis location, and the patient environment or location are all potential variables that may affect an outcome. These factors can, and should, influence the design of the research protocol.

Most interestingly, the science of informatics, coupled with the large databases in pathology information systems, has provided new research opportunities. The area of most interest to us at the University of Alabama at Birmingham is the area of data mining and what we term *knowledge discovery*. The hypothesis is that buried within the large and complex pathology databases are patterns of information that have been heretofore unrecognized and that represent new information relevant to appropriate and effective health care delivery. If this is true, then application of data mining to pathology databases may have an impact on medicine that rivals the impact of many previous key technological advances in pathology, such as automation, immunoassays, and molecular diagnostic techniques. The remainder of this report is devoted to this topic.

KNOWLEDGE DISCOVERY AND DATA MINING

The Challenge

As the body of data we collect grows in both size and complexity, we must explore new methodologies to analyze it. Our ability to collect and store data has grown proportionally faster than our ability to analyze data. It is estimated that the world's information load doubles every 20 months, with the number of databases growing at least as fast. As a result, there is a general consensus among experts that significant untapped knowledge lies hidden in many large databases.

Knowledge discovery and data mining (KDDM) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.[6] The challenge of KDDM is to make effective use of large databases to discover the untapped knowledge that lies hidden therein. Specifically, it is the task of developing and implementing methodologies that can discover interesting, useful patterns and concepts in databases that can be used (1) to support mission-critical decision making, (2) for prediction and classification tasks, (3) for summarization, (4) for surveillance, or (5) to explain observed phenomenon.

Some candidate applications for KDDM are shown in Table 3. One example of this process as applied to the retail industry, the purchase of diapers at a convenience store, serves to illustrate the utility of this approach. A large convenience store chain asked that the national database of their sales records be mined for interesting associations. This resulted in the detection of a peak in the purchase of disposable diapers in the afternoon, coincident with individuals leaving work. The hypothesis was that individuals were buying these diapers on their way home from work, and the merchant hypothesized if they arranged the diapers next to snacks and beer, the sales of the latter items would increase. As predicted, the sale of

Table 3.—Potential Applications for KDDM in Health Care

- I. Health care
  - A. Surveillance
    - 1. Outcomes
    - 2. Epidemiology
  - B. Clinical practice guidelines/critical pathway evaluation
    - 1. Identify important decision variables
    - 2. Evaluate efficiency of current guidelines
    - 3. Predictive modeling
    - 4. Deriving rule-based systems
  - C. Imaging
    - 1. Classification
  - D. Molecular sequence analysis

both snacks and beer increased dramatically during that time.

Traditional methods of data analysis are manual and confirmatory in nature. Database queries are manual in that they must be formulated by a user. The results of the query are often used after some manipulation to answer specific questions. However, manual queries are limited in that they are formulated with specific questions in mind and are used in answering those specific questions.

The issue is further complicated because traditional (confirmatory) statistics are based on hypothesis testing. Hypotheses are statements about the data that are either rejected or not rejected based on statistical testing. Therefore, traditional statistics offer us methods to confirm or reject a hypothesis. However, traditional statistics cannot lead to discovering patterns that we do not already suspect. In other words, in traditional statistics, the search for useful relationships in the data (knowledge) is based on the expectations of those generating hypotheses. Results, therefore, are the verifications, or lack thereof, of the suspicions of the investigators. These methodologies do not offer us a way to discover hidden patterns in data, resulting in answers only to the questions that are asked.

Knowledge discovery and data mining is an emerging, interdisciplinary research field that lives at the intersection of the computer science (database, artificial intelligence, graphics, and visualization), statistics, and several application domains. Knowledge discovery and data mining is a process (traditionally called the KDD[6] Process) that is composed of a number of steps. Briefly, these steps are as follows:

- Develop an understanding of the application domain.
- Create a target data set on which discovery is going to be performed.
- Clean and preprocess the data.
- Reduce the data.
- Choose a data mining task (decide whether the goal is classification, regression, clustering, description, modeling, etc).
- Choose a data mining algorithm.
- Mine the data; ie, search for patterns.
- Interpret mined patterns.
- Consolidate the discovered knowledge (incorporate it into decision systems, reports, etc).

The shortcoming of traditional data analysis on large databases is that one does not know what questions to ask (hypotheses to test) to discover hidden knowledge. Knowledge discovery and data mining addresses that shortcoming by discovering interesting patterns that the user does not suspect. In fact, one can view the KDDM process as one that generates, or suggests, testable hypotheses[7] that can be independently tested via traditional confirmatory methods.

Many of the steps of the data-mining process are determined on an application-by-application basis and can be quite involved, and each has open-ended research issues. The KDDM research group at the University of Alabama at Birmingham is actively addressing many research issues as it is developing, building, and testing a KDDM system. This system, the first test of which focuses in the general domain of microbiology databases, represents to our knowledge the first application of the KDDM process to pathology databases. More specifically, the initial testbed for this powerful software system is a large antibiotic susceptibility testing database. New patterns of information are being uncovered, and the significance of these patterns is being tested.

Because health care and the medical sciences are certainly no stranger to large databases, and because the amount and complexity of data are increasing rapidly, KDDM should have a place in the future of medical information research. The amount of undiscovered knowledge in medical databases is potentially very large. As health services and outcomes research becomes more important and as more data are collected, there will be a need for new methodologies to discover hidden knowledge. These new methodologies are likely to make a tremendous impact on the future of health care around the world.

It is this kind of research, in addition to more standard and routine uses of pathology information systems, that will have impact on the health care delivery system of the future. The pathology databases represent an outstanding resource, which we as pathologists and clinical laboratory scientists must use routinely in such areas as outcomes analysis, cost-effectiveness analysis, and guideline implementation. Possibly more exciting is the opportunity to use our databases to make an even broader impact on health care via the application information sciences using techniques such as KDDM.

This work was supported by cooperative agreement U47/CCU411451 (Dr Moser) and National Institutes of Standards and Technology, Gaithersburg, Md, grant subcontract 94-04-008 with CERNER Corporation (Dr McDonald).

References

1. Buaki GJ, Moreau DR. Laboratory computing—process and information management supporting high-quality, cost-effective healthcare. *Clin Chem.* 1995; 41:1338–1344.

2. Connelly DP, Sielaff BH, Willard KE. A clinician's workstation for improving laboratory use: integrated display of laboratory results. *Am J Clin Pathol.* 1995; 104:243–252.

3. McDonald JM, Smith JA. Value-added laboratory medicine in an era of managed care. *Clin Chem.* 1995;41:1256–1262.

4. Feussner JR. Evidence-based medicine: new priority for an old paradigm. *J Bone Miner Res.* 1996;11:877–882. Editorial.

5. McDonald JM, Friedberg RC, Moser SA, Smith JA. Role of the laboratory professional–new opportunities. In: 1995 Institute on Critical Issues in Health Laboratory Practice, Frontiers in Laboratory Practice Research. Atlanta: Ga: Centers for Disease Control and Prevention; 1996:335–341.

6. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery; an overview. In: Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, eds. *Advances in Knowledge Discovery and Data Mining.* Menlo Park, Calif: AIII Press/MIT Press; 1996:1–34.

7. Fayyad U, Piatetsky-Sharpiro G, Smyth P. From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine.* 1996;17:37–54.