# The National Library of Medicine and Medical Informatics

DONALD A. B. LINDBERG, MD, *and* HAROLD M. SCHOOLMAN, MD, *Bethesda, Maryland*

*Medical informatics attempts to provide the theoretic and scientific basis for the use of automated information systems in biomedicine. Even though a new field, its roots are in the 19th century. The National Library of Medicine (NLM) began classifying the medical literature and publishing the* Index Medicus *in 1897; in the early 1960s, the growth of the index gave rise to MEDLARS, the first successful, large-scale, computerized bibliographic system. In 1971, about the time MEDLARS evolved into a nationwide on-line retrieval system known as MEDLINE, a committee of the Association of American Medical Colleges published a report calling for the NLM to exert strong leadership in developing computer applications for information transfer in medicine. The NLM has sponsored several training and research programs in this area and is now developing the concept of "centers of excellence" in medical informatics. In addition, there are a number of current research and development activities within the NLM internal and extramural programs that may influence the progress of medical informatics.*

(Lindberg DAB, Schoolman HM: The National Library of Medicine and medical informatics, *In* Medical informatics [Special Issue]. West J Med 1986 Dec; 145:786-790)

Medical informatics, a new, multidisciplinary field, attempts to provide the theoretic and scientific basis for the use of automated information systems in biomedicine. Investigators come from health and computer sciences, engineering, library and information science and management

Like many new fields, especially in medicine, the initial practical demonstrations of the power of the methods of medical informatics (in this case, to store and retrieve information, to compute quantities and to control events) precede and overshadow its incomplete theoretic base. Nonetheless, this field is already having a profound effect on health care affairs because of its focus on the fundamental processes of human medical treatment and information. Thus, studying the processes of teaching and learning the accumulated record of scientific evidence and discovery, the general problems of patient record-keeping and the processes of medical diagnostic and management decision-making has been of greater interest to those in medical informatics than issues particular to the individual specialty areas of medicine and health care per se.

For example, the use by a surgeon of information from statistical studies of populations along with incomplete sets of laboratory, roentgenologic, scientific and clinical information about a single patient to form the basis for the surgeon's decision to operate has been of great interest in medical informatics. Sohn, Robbins[1] and de Dombal[2] have reported the results of such studies. These are examples of the general clinical problem of making decisions in the face of uncertainty. Such a dilemma has since the beginning of medicine placed a premium on learning, on skills in reasoning and on empathy. Since the beginning of modern scientific medicine, there has been a premium, too, on the thoroughness of a physician's knowledge and his or her familiarity with the rapidly changing understanding of disease processes and the discovery of new treatments and preventatives.

The National Library of Medicine (NLM) has for 150 years been dedicated to organizing the knowledge of medicine and to providing timely information to health professionals in support of clinical health care and research. More recently, NLM has also taken a strong interest in the cognitive processes and systems with which such knowledge is used in timely and appropriate ways to improve health. Some aspects of this evolution may be of interest to readers.

The National Library of Medicine has the legislative mandate

to assist the advancement of medical and related sciences, and to aid in the

From the National Library of Medicine, Public Health Service, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland.

Reprint requests to Donald A. B. Lindberg, MD, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894.

ABBREVIATIONS USED IN TEXT

AAMC = Association of American Medical Colleges
GRACE = Graphic Arts Composing Equipment
MEDLARS = Medical Literature Analysis and Retrieval System
MeSH = Medical Subject Headings
NLM = National Library of Medicine

dissemination and exchange of scientific and other information important to the progress of medicine and to the public health.[3]

## Early NLM Development

Indeed, it was in the pursuit of this purpose that John Shaw Billings, MD, in 1879 started the *Index Medicus*, "A monthly classified record of the Current Medical Literature of the World." In his prospectus, Billings described the new index as follows:

In its pages the practitioner will find the titles of parallels for his anomalous cases, accounts of new remedies, and the latest methods in therapeutics. The teacher will observe what is being written or taught by the masters of his art in all countries. The author will be enabled to add the latest views and cases to his forthcoming work, or to discover where he has been anticipated by other writers, and the publishers of medical books and periodicals must necessarily profit by the publicity given to their productions.[4]

The first issue of *Index Medicus* listed 18,000 articles. Within 50 years its annual coverage had tripled. Today the annual coverage exceeds 250,000 articles. By the early 1950s the logistics of maintaining the *Index Medicus* had become overwhelming. Currency could not be maintained and backlogs were growing rapidly.

In 1960 a new system was put into operation that handled 125,000 citations with multiple entries for each. The system used tape-actuated typewriters to repetitively print a unit record across the top of a Hollerith card, the keypunching of filing indicia into a reserved area of each card, sorters and collaters to arrange the cards into appropriate author and subject subsets and a high-speed rotary-step camera. The camera, capable of varying its aperture to fit the number of lines in an entry, would photograph the cards on filmstrips that could be cut into column length.

### *The Creation of MEDLARS*

This early application of "automation" foreshadowed what might be done and led to the specifications of the Medical Literature Analysis and Retrieval System (MEDLARS). It was this system that produced the January 1964 issue of *Index Medicus*. A machine called GRACE (Graphic Arts Composing Equipment), developed specifically for the NLM, typeset pages at the rate of 300 characters per second. GRACE accepted input directly from magnetic tape that had been reworked by a computer into a page-block format.

This photocopier tape contained a matrix of 226 characters etched on glass. These characters were in several sizes and fonts, with a complement of diacritical marks. Behind each character on the matrix was a high-speed flash tube; the circuitry of GRACE timed the flashing of these lights. Between the matrix plate and a 9-in wide roll of film, there was a mirror and reciprocating lens that constantly roved back and forth, photographing one line of characters, character by character, across the entire width of a three-column page. It may be calculated that, at 1.7 seconds per sweep, GRACE had to labor for 150 hours to compose the five volumes and 8,900 pages of the 1968 *Cumulated Index Medicus*.

Remarkable for its day, GRACE was a significant step in computer-controlled typesetting. By showing that such a system was operationally feasible, NLM created the precursor, if not the stimulus, for computer-based publishing. This was recognized by GRACE's retirement in 1969 to the collection of the Smithsonian Institution.

But MEDLARS was not designed solely for the publication of *Index Medicus*. From the beginning, retrieval was prominent in its specifications. By the late 1960s, around 16,000 searches a year were being done in batch mode on three computers that mounted the MEDLARS tapes. MeSH (Medical Subject Headings) had been thoroughly revised and greatly expanded. This structured, controlled vocabulary permitted in a computerized search a far greater specificity than was possible through the use of the *Index Medicus* main headings.

The newly formed Lister Hill Center soon began experiments with interactive on-line searching using the teletypewriter communication network (TWX) and an abbreviated file of 100 clinical journals called Abridged Index Medicus—thus the name AIM-TWX. The extension to the full *Index Medicus* file and the utilization of common communication carriers soon followed, and MEDLINE was launched. The immediate and remarkable success of MEDLINE provided a great stimulus to the common carriers to rapidly expand the coverage of their telecommunication networks. This contributed greatly to the early introduction of on-line searching of many other subsequent data bases.

The on-line interactive search made the rapid identification of citations a practical reality. But citations were not the answer to a searcher's question. For that one needed the documents to which the citations referred. To assure the availability of documents, NLM, under the aegis of the Medical Library Assistance Act of 1965, supported new library construction, collection development, training and special publications.

Perhaps most important of all, however, the Act extended to NLM authority to establish a national communications system, including, if necessary, construction of NLM branch libraries in the states and construction or reconstruction of libraries at the medical schools. Fortunately, this authority was used wisely to build up strong local and regional libraries and to build the Regional Medical Library network that we know today.

In this system, documents that were not available at a hospital library could be requested from larger resource libraries and a Regional Medical Library. If the material sought was not available there, NLM acted as the library of last resort. In 1985 more than 2 million documents were delivered within the network. NLM had to supply only 160,000 of these. This network-routing system is being even more fully automated. A national computer-based system for interlibrary loan requests (DOCLINE) is now being implemented in stages throughout the nation.

## Medical Education

While the Lister Hill Center was beginning to engage in experiments in telemedicine supporting health care delivery and medical education using the Advanced Technology Satellites ATS-1 and ATS-6, the NLM contracted with the Association of American Medical Colleges (AAMC) to assess the

future role of computer and communications technology in medical education and health care delivery. The report of the AAMC committee, chaired by Eugene Stead, MD, was published in 1971. The committee wrote with regard to computers in medicine

> that the time has come for the Lister Hill Center to exert strong leadership in the development of the computer science field as it relates to medical education and the preparation of professionals for the delivery of health care. . . . The committee believes that we will have to reexamine the problems in medical education and practice and devise solutions utilizing computer technology which will offer new approaches to education and give new patterns for professional staffing of clinical units, for the collection of clinical and laboratory data, and for clinical decision making.[5]

These recommendations led to a revamping of the intramural programs of the Lister Hill Center and to the establishment of a series of research, training and career development grants in what was then called "Computers in Medicine."

In 1971, the vast majority of academic health centers barely tolerated such endeavors. There was no locus for such activity, and, indeed, there was little recognition that such a field existed. It certainly had no name. Not even the people working in the field could agree on what it should be called, let alone what it encompassed. Years later, the name "medical informatics" was adopted.

For these reasons the training grants had as their primary stated objective the training of established faculty members who could return to their institutions not so much to research careers in medical informatics—although that, of course, was not discouraged—but rather to improve the environment and to help create a greater institutional receptivity for the increasing use of automated systems in all phases of medical education and health care delivery. The research and career development grants, on the other hand, were designed to support and enhance the careers of those few existing research workers in the field.

The extent to which the training grants accomplished their avowed purpose is difficult to assess definitively. These efforts, the emergence into prominence of general information systems and microcomputers and the rapid introduction into medicine of computer-mediated technology such as the computed tomographic scan all contributed to a radical change in environment in the decade that followed. Because of that change, the training grants were changed in 1982 to research training grants with the specific purpose of training future researchers in the field.

Related to this, the NLM has recently proposed the development of centers of excellence in medical informatics.

### *Centers of Excellence*

Don E. Detmer, MD, Vice President for Health Sciences, University of Utah School of Medicine (Salt Lake City), views the center of excellence as "an institutional locus for the education of health professionals and medical informatics experts in the techniques and problems of using computers and information systems to improve health care."[6] Jean Mayer, PhD, President of Tufts University School of Medicine (Boston), in proposing centers of excellence in medical informatics describes them as providing "a highly productive research and scholarly environment with the status of a division or department in a medical school."[6]

However it is described, this organization would have three major functions: coordinating medical informatics studies and skills into the overall curriculum; conducting research in medical informatics, and training and educating in the field. Such centers of excellence will substantially enhance the recognition and career development so badly needed in the field.

Yet, it is not only the environment for medical informatics that has been changing dramatically. The very boundaries of clinical medicine are rapidly being pushed back to the molecular level, and with that development has come an explosion of automated data bases. As the volume of medical information has rapidly expanded, access to that information has become more and more constrained by the system that housed it. The utility of information to a very large extent depends on the ease with which it can be retrieved. Two great bottlenecks to information retrieval are the difficulty of human-to-machine and machine-to-machine communication.

## Knowledge Representation and Medical Informatics

In the late 1950s, the modern NLM classification system, MeSH, was rebuilt. It combined through a common set of pointers the indexing of articles and cataloguing of books and monographs. This controlled, highly structured, knowledge representation scheme is a very powerful tool for retrieving citations to the medical literature. Its structure and "explode" capability make it possible through single entries to retrieve a wide range of related material. The in-depth indexing (an average of 12 terms assigned per article) and the use of Boolean manipulators ("and," "or" and "and not") make possible precise retrieval. But the system is complicated and was originally designed for relatively well-trained users. Indeed, at first the training period for MEDLARS searchers was three months. While this has been greatly shortened, the vast majority of searches are still done by trained intermediaries.

### *GRATEFUL MED*

Numerous attempts have been made by NLM and others to simplify this machine interface so that the system may be used by nonexperts. For certain types of searches these systems are very effective. NLM's user-friendly connection to MEDLINE, called "GRATEFUL MED," provides a transparent, menu-driven use of Boolean operators "and" and "or" with free language entries that are searched both as free text and as MeSH-controlled vocabulary.[7] Even in its simplest form, using only selected features of the power of the MEDLARS system, it will retrieve enough relevant citations to satisfy many queries. On the other hand, more experienced users can operate through another mode of GRATEFUL MED to bypass the menus and have all the power of MEDLARS command language searching restored. We believe GRATEFUL MED will also be helpful to persons who wish to learn more about information-retrieval issues and to improve their skills in searching through a gradual and self-controlled tutorial process. Successive versions of this front-end search system are already being tested. Version 2.0 incorporates much of the Medical Subject Headings thesaurus of index terms. This is designed to assist biomedical inquirers to "map" their idiosyncratic or specialists' inquiries to the (often-more-general) MeSH terminology to obtain access to the larger number of scientific articles that are indexed under standard index terms.

GRATEFUL MED already includes a simplistic algorithm that suggests possible improvement in phrasing a user's search statement. It seems likely that future versions will offer advice to users that is an increasingly intelligent equivalent of the assistance of an expert reference librarian. (GRATEFUL MED is available for $29.95 [1986 price] from the National Technical Information Service, United States Department of Commerce, Springfield, VA 22161; order number PB86-158482.)

*MICRO-CSIN*

An obvious improvement in user-friendly front ends is for them to be able to communicate with more than one data base. MICRO-CSIN (Chemical Substances Information Network) is such a system. It knows the protocols and can, therefore, automatically log into any of the hundreds of data bases available from nine main vendors. Because its basic purpose, as the name implies, is to search chemical systems, cross-data-base searching is reasonably effective. This machine-to-machine communication is facilitated by the presence of unique identifiers in chemistry such as the chemical name and the Chemical Abstracts Service registry number. These unique identifiers make it possible to find the chemical of concern regardless of the data base and the data-base structure. A small number of key concepts, such as toxicity, physical and chemical properties and so forth then make it possible to retrieve the record of interest. This system, too, offers many opportunities for improvement through adapting the growing understanding of informatics principles. Systems such as MICRO-CSIN do not, for example, really "know" a user, much less his or her educational level, state of urgency in making the request or basis in having (perhaps recently) received other information from the system. Such possibilities for systems to do "user profiling" are still only research possibilities, even though the fundamental concepts involved are easily understandable at the common-sense level.

*Unified Medical Language System*

Unfortunately for most of biomedicine, there are no unique identifiers. Even where data bases have a shared subject domain, they use different knowledge representation schemes, different indexing philosophies and different vocabularies. In these instances, transferring a concept from one data base to another is extremely uncertain, being subject to all the vagaries of normal discourse without a human to interpret them. There are three possible solutions to this problem: adopt a standard vocabulary, interpose an "expert" system that simulates a human intermediary or create some canonical knowledge representation scheme into and from which all others can be transposed.

The NLM believes the first to be unacceptable to the community. A combination of the second and third is being studied with the help of the academic community and professional societies. This combination has been called a "Unified Medical Language System." It is viewed as a long-term (and costly) project from which many intermediary benefits may be derived in the course of its pursuit. The NLM has received congressional support to coordinate this effort. Resources from five universities have been contracted for, and a significant NLM staff effort has been started. The American Medical Association has agreed to coordinate the involvement of professional societies.

The rate of increase in biomedical knowledge makes it impossible for any one person to be truly up to date. One of the greatest challenges of the information age is to create effective syntheses that can be maintained automatically. The traditional medical syntheses of books and monographs are now almost always substantially out of date before they are printed. Review articles are somewhat better, but in the absence of some updating mechanism they, too, are soon out of date.

*Hepatitis Knowledge Base*

The National Library of Medicine in 1979 built a prototype "hepatitis knowledge base" aimed at providing up-to-date information needed to support daily diagnostic, prognostic and management decisions.[8] The hepatitis knowledge base was designed to provide substantive information relevant to a wide variety of questions rather than citations to the literature that may contain such information. The information in the knowledge base was hierarchically arranged to provide increasing detail and specificity, including citations to the primary publications on which the syntheses were built. For this limited domain, the structure of the hepatitis knowledge base was, indeed, a knowledge representation scheme.

*Expert Systems*

In a related area, the NLM has supported intramurally and by grant the development of knowledge-based "expert systems." They have in common the components of a knowledge base and an "inference engine." To a varying extent, each expert system may model human cognitive processes in such areas as medical diagnosis and patient management. In all these cases the knowledge base is constructed on some representation scheme (often sets of production rules) and represents a kind of synthesis for that domain. Updating and maintaining such syntheses remain important problems for which no adequate machine solution has yet been devised.

*Image Processing*

In the foregoing discussion of NLM and medical informatics, we have attempted to suggest the progress of the solutions to the problem of providing health scientists with the information they need when they need it and in a format that can be readily used. We have, however, dealt only with information represented by symbols—that is, words, numbers and so forth. Much medical information resides not in symbols of language but in images. Therefore, NLM also supports research in image processing.

The traditional problems of storage, indexing and retrieval are in the case of images compounded by problems such as color fidelity and resolution. Modern technology such as optical disc storage and digitization of images suggests enormous possibilities in the effective use of information from images. Perhaps the main bottleneck resides in indexing. Verbal labeling of images sacrifices an enormous amount of information. This makes retrieval very inefficient except in its simplest form. One can visualize retrieving information from and across images based on spatial clusters in a digitized format. Indeed, the digitization of the images makes possible the three-dimensional projections that conform to reality. Because much of the imaging of modern medicine—computed tomographic scans, nuclear magnetic resonance and the like—is digital from the beginning, this description is not a speculation about the future but rather a view of the present.

## Institutional Settings

But solving the technical problems of knowledge representation, unified medical language systems and machine-to-machine communication is not all that is required. Institutional restructuring to manage effectively the complex information activities within health centers must also occur. For that reason, NLM instituted in 1983 a program to support the study, planning and effectuation of Integrated Academic Information Management Systems.[9] This program was initiated with planning grants and contracts at four academic health centers; four additional awards were made a year later. Experimentation is now going on and several models will be implemented in the next few years.

NLM, in collaboration with the National Institutes of Health Division of Research Resources, has supported research in medical informatics to make possible the rapid application at the bedside of the advances in medicine resulting from research. Now, at a rapidly increasing rate, the view of biology is shifting to the molecular structure of genes and their protein products. Research in the life sciences is becoming increasingly dependent on tools to store and manipulate large amounts of data on the behavior and structure of macromolecules. The ability to measure and change events occurring on a molecular level is particularly significant where the development of techniques to sequence, clone and remodel genetic material is leading to the control of life processes with a precision never before known. The continued pursuit of this knowledge is as much an information-processing problem as it is a problem in biology. Unfortunately, the institutional infrastructure for the support of the information-processing part of these endeavors is far less defined and developed than that for the biology. NLM already plays a crucial role in its bibliographic control of the published literature. The management of the rapidly emerging banks of data needs now to be organized and integrated.

With such important issues facing it, the NLM is now engaged in a careful look to the future. The initial phase of a long-range planning activity has just been completed in which NLM has had the views and advice of five panels of distinguished outside experts. The topic, "Research in Medical Informatics," was carefully and specifically considered by one of the panels.

The final recommendations of the overall panel are now under review by our board of regents. It is already clear, however, that the board readily accepts the importance of automated information systems and knowledge bases in its vision of the future and that research in medical informatics is critical in assuring the best of these future developments.

### REFERENCES

1. Sohn N, Robbins RD: Computer-assisted surgery (Letter). N Engl J Med 1985 Apr; 312:924

2. De Dombal FT: Computers and decision-making: An overview for gastro-enterologists, *In* Rosen P, de Dombal FT (Eds): Computer Aid in Gastroenterology. Basel, Karger, 1984, pp 119-133

3. National Library of Medicine Act of 1956. PL 941 (84th Cong, chap 907, 2nd sess) §3430

4. Billings JS: Prospectus. Index Medicus 1879 Jan; 1:1

5. Stead EA Jr, Smythe CM, Gunn CG, et al: Educational technology for medicine: Roles for the Lister Hill Center. J Med Educ 1971; 46:(pt 2):59-66

6. Departments of Labor, Health and Human Services, Education and Related Agencies Appropriations for 1987: Hearings before a Subcommittee of the Committee on Appropriations, House of Representatives, 99th Congress. Government Printing Office, 1986

7. Abrutyn E: Software: Grateful Med 1.0. Ann Intern Med 1986; 105:321

8. Bernstein LM, Siegel ER, Goldstein CM: The hepatitis knowledge base: A prototype information transfer system. Ann Intern Med 1980; 93 (pt 2):169-181

9. Matheson NW, Cooper JAD: Academic information in the academic health sciences center—Roles for the library in information management. J Med Educ 1982; 57(pt 2):1-93