BIOMEDICAL INFORMATICS: PRECIOUS SCIENTIFIC RESOURCE AND PUBLIC POLICY DILEMMA

DONALD A. B. LINDBERG

BETHESDA, MARYLAND

ABSTRACT

Biomedical informatics includes the application of computers, information networks and systems, and a growing body of scientific understanding to a range of problems. As skill in this field increases and as progress in virtually all modern biomedical science becomes more data intensive, informatics becomes a precious resource. Applications areas include access to knowledge, discovery in genomics, medical records, mathematical modeling, and bioengineering.

At the same time, progress in informatics is deeply dependent on resolution of four major public policy issues: digital intellectual property rights, genetic testing protection, medical data privacy, and the role of biomedical data in the context of information warfare and homeland security.

INTRODUCTION

The longstanding biomedical interest in "computers in medicine" (1), computer based patient record systems (2), computational linguistics (3), "supercomputers" (4), and the Internet and its successor Next Generation Internet (5) are focused nowadays in the field of biomedical informatics. This work has high relevance to a number of scientific and medical efforts. At the same time, resolution of some major public policy issues in the U.S. stand as obstacles to full utilization of the new techniques.

MATERIALS AND METHODS

This presentation will be based upon publicly available records and products of the U.S. National Library of Medicine, the Office of High Performance Computing and Communication of the Executive Office of the President, and bills and hearings in the U.S. Congress.

From the National Library of Medicine, Bethesda, Maryland

RESULTS

Key Application Areas

I. Appropriate access by individuals to medical knowledge and understanding

NLM has produced the paper-based *Index Medicus* since 1879, the computer printed version since 1964, its electronic searchable version MEDLINE since 1971. Distribution of this index to the biomedical scientific literature has improved along with the public telephone, value added electronic networks, and Internet. At all times NLM insisted on supporting electronic access and a pricing scheme that made access to our files equally available at identical charges throughout the entire U.S. This has meant that doctors and patients can count on access to the latest medical information no matter where they live. The number scientific journals included has increased as well, from around 100 in 1971 to about 4500 in 2002.

Indexed plus factual data in toxicology and environmental health was added in 1976 in response to the publication of *Silent Spring* by Rachel Carson and the resultant Congressional authorizing (and regulatory) legislation. Compatible index information in bioethics was added in 1978.

Like all good libraries, NLM has always permitted use of any part of its collection or services by any person. Yet until 1996 NLM had operated under the mandate (of its Board of Regents, the House Appropriations Committee, and OMB) that we serve the public primarily via service to scientists and health care professionals, and this accompanied by charges for the marginal costs of the information services. On April 16, 1996, Senator (and surgeon) William Frist made the first Grateful Med search of MEDLINE using Internet. This new method of telecommunication, when combined with the software innovation called the World Wide Web, resulted in a reduction in our actual costs of providing MEDLINE searching by 93%. On June 26, 1997, Senators Arlen Spector and Tom Harkin hosted a public demonstration of the first completely free MEDLINE search (by using the new NLM software PubMed); the actual search was performed by the Vice President, Albert Gore. Subsequently the volume of searches of MEDLINE increased remarkably. Even more surprising to us, the percentage of searches by the public increased from about 1% to 30%. Taking into account this major new group of users of our medical and scientific data base MEDLINE, we launched a new data base called MEDLINEplus beginning on October 22, 1998. The number of health care topics covered grew from the initial 20 to 577 by October 2002. Along with elimination of user charges, since 1997 searching has been performed without requiring that users provide us any form of personal identification. Usage has grown dramatically, so that the total searches per day for MEDLINEplus, MEDLINE, and the GenBank files now measures more than one million.

II. New knowledge in Genomics

NLM was authorized by Congress in 1988 to create the National Center for Biotechnology Information (6). In 1992 NCBI began operating GenBank as a means to organize, store, and share DNA sequences from both large scale sequencing efforts such as the Human Genome Project as well as from individual laboratories throughout the world. Since 1992, the database has grown from 71,000 to 20 million sequences. GenBank operates as part of an international collaboration, sharing data daily with its European partner, the European Bioinformatics Institute in Hinxton, UK, and with its Asian partner, the DNA Database of Japan in Mishima, Japan. The database contains over 6 million human DNA sequences that include nucleotide base pairs as well as the sequence data from over 136,000 other species.

Understanding of genomics has grown vastly as well. Expressed Sequence Tags, Single Nucleotide Polymorphisms and the new Haplotype Maps thereof, the production of many proteins from a single gene, the existence of many non-coding cDNA's, and the seemingly endless number of up and down regulators of protein production—these are all examples of understanding of genomics that derived from free, online access to vast amounts of validated (and curated) experimental data from many species retained in the public domain. It's apparent that "information retrieval" within files of many billions of items would be meaningless without the embedded metadata and the associated software to display or operate on the information so as to assist the user. In many cases the major contribution of informatics has been visualization tools to help scientists see new relationships in genetic data. These have already resulted in calls for a molecular taxonomy of pathological states, of (ambitious) plans to replace symptomatic descriptors of patient groups with meaningful genetic stratifiers (e.g. drug responders vs non-responders).

III. Mathematical Modeling and Understanding of Human Anatomy

NLM commenced the Visible Human project in 1991 to produce a computer-based digital representation at a submillimeter resolution of an entire adult male and a female body, retaining all the 3-dimensional information including fiduciary marks relating cryosection, CAT scan, and MRI representations. Since then more than 1700 (free) licenses to use these data have been signed by educational institution, individuals, and firms in 45 countries. Scores of applications have been produced. These include implementation of computer based systems to train clinicians in medical procedure, especially invasive procedures such as bronchoscopy or regional nerve blocks for anesthesia (7,8).

IV. Computer-based health records

Forming search requests to an electronic system as well as choosing terms to record a patient's complaints, findings, or clinical course both demand use of a precise, meaningful, sufficient, and controlled vocabulary. All other features of these systems are mere window dressing. Hence NLM began in 1985 to build the Unified Medical Language System. The UMLS was designed to provide individual users and hospital information system builders access to many existing biomedical thesauri and a system that understands the medical meaning of the individual terms and their interrelationships. Subsequently the system has had extensive use in practice and research settings (9,10).

V. Bioengineering developments

In addition to the important traditional bioengineering functions in instrumentation and biomaterials, many new systems are making fine use of informatics software. Magnetic Resonance Imaging (itself computing intense) is now being complemented by Functional MRI applications that combine physical imaging of the brain with registered images that reflect physiological functions synchronized with tasks and external stimuli.

Implantable cardiac pacemakers also are now being complemented by new programmable "intelligent" pacemaker/defibrillators. These remarkable developments have exceeded all expectations by actually improving cardiac output in failing hearts (11).

DISCUSSION

The success of the informatics developments cited above, plus many not cited, and the rapid advances of relevant biomedical research have moved faster than the pace of national consensus about a number of public policy matters. Five examples are noted below.

I. Intellectual property rights and public access to information

In the case of NLM's use of authors' summaries, along with bibliographic citation data, to announce and index biomedical scientific publications, the federal courts judged the practice proper—in spite of numerous objections by commercial publishers (12). In the case of linking to Internet information in MEDLINEplus, NLM selects information already within the public domain or leases access to it from copyright holders in order to present it to patients, families, and the public (13). Many commercial companies vend consumer health information via the Internet, or offer it without charge while selling products to the user or selling the identity of the users to commercial sponsors. Thus far the courts have not questioned our practice nor the commercial practice. Nonetheless NLM does not link to information sources (even good ones) if the user is required to give personal identity in exchange for access to the information.

In the case of genomic information, public—or even governmental—policy is not wholly firm. All data from publicly supported Human Genome Project sites was deposited in raw form immediately and daily, and was incorporated immediately into GenBank at NLM/NCBI and the other two international data bases. The Bermuda Agreement among the participating international sequencing centers and funders made this explicit (14). Sequencing centers funded by venture capital generally kept their data private, charged for access, or imposed "reach through" rights on future developments based on seeing their data (14).

The practice and policy concerning publicly funded human and other genomic data are variable. Some institutions require immediate release to the public; some tolerate delays of months or even years. Generally the former work is funded as competitive research contracts; generally the latter is funded as research grants. The argument for immediate release is partly the fundamental justification of spending public funds in order that discoveries enter the public domain; partly it is the amazing usefulness to the broad community of scientists of genomic data even in raw form. The argument for delay is that sequencers (and presumably other genomic investigators) ought to have a decent chance to "publish first" on data coming from their labs no matter how the work was financed. Lack of resolution of this difference in views is most extreme at the moment in the case of data on structure and function of the protein products of genes.

II. Medical data privacy Anatomy

PL 104-191, the Health Insurance Portability and Accountability Act of 1996, provided basic protections for medical data privacy in computerized health records. It reserved to the Congress to legislate details over the next three years, failing which the Secretary of HHS was to issue regulations on the matter having the force of law. A variety of studies of the privacy needs were done (15). The Office of the Secretary issued many notices and drafts of the privacy regulations. An amazingly large number of comments were recorded; indeed one version drew 54,000 written responses. Some concerns derived from anticipated practical difficulties, e.g. could an individual pick up his or her spouse's prescription at the (perhaps remote) pharmacy without written permission? Some objections stemmed from principle, e.g. would the federal regulations limit benefits already provided by a state? Ultimately the Secretary issued the regulations. Doubts about the actual difficulty in complying with HIPAA delay development and testing of a number of potential informatics advances. These include testing and use in the U.S. of "smart cards" for personal identification, authentication, and for bearing actual health data. One notes that all three uses (plus commerce) are widespread in Europe already. Hospital managers cite many other operational difficulties with implementing HIPAA—even though there is almost universal agreement that patients fully deserve to hold private their medical records when they wish, and wide agreement that true dangers of violation of privacy exist (16,17). Absent the hoped-for legislative solution and the reliance on a long and complex regulatory process, a number of regrettable situations occurred, for example the absence of security standards preceding the issuance of the privacy regulations. This naturally confuses and delays information system builders and the development and testing of biosensor technology.

III. Genetic testing rights

Here too there is almost universal agreement that a person ought not be denied employment, insurance, marriage license, and most ordinary social benefits simply because of public access to the results of a genetic test—especially if the test had been somehow compelled of the individual—or perhaps if the test resulted from voluntary participation in a clinical trial. Legislation introduced in the 107th Congress would have barred employers from requiring genetic testing as a condition for employment, prevented health insurers from denying coverage based on test results, and prevented disclosure of genetic information to third parties. Yet no attempt to pass federal legislation with these goals in mind has succeeded. Absent federal legislation that provides adequate protection against discrimination, there will be substantial reluctance to take advantage of these new capabilities for combating disease.

There is little doubt among clinical trials investigators—especially those in genomic work—that the lack of such protection is already seriously impeding recruiting subjects. A lamentable secondary consequence of fear about genetic wrongs was the unjustified arguments amidst the debates on provision of medical data privacy that all genetic information should be excluded from the electronic medical records and made subject to special secrecy procedures. Actually genomic insights seem to be informing virtually every medical specialty, so that it was finally apparent that data privacy and protection against misuse of genetic data were really two separate issues.

IV. Information warfare/homeland defense

Concern with the need to prepare against bioterrorism has understandably taken front place amongst the requirements put upon biomedical researchers (18). Yet other public policy matters still to be enunciated will impact future biomedical informatics contributions to scientific progress. For example, what are to be the requirements upon hospital computer security in the context of information warfare? Currently NLM must support five full time computer technicians solely to protect our computers and (more importantly) our communication networks and files against continual attempts to "crash" or deny access to the systems. In February 2002 there were 20,000 such attacks—almost all defeated. Imagine such nonsense when we have no classified data and no individual patient records! The only gratification to such attackers is that from destroying and vandalizing. From our point of view, of course, "only" vandalizing can destroy systems whose re-creation can easily cost many tens of thousands of dollars. Would things become simpler and easier for hospitals with computer-based patient records were the terrorism (rather than seeming vandalism) to increase? Already a number of legislative bills have been offered in the Congress aiming to protect against such abuses of the communication networks.

In conclusion, the computer and communications-based techniques introduced or preempted by biomedical informatics research and development are powerful complements to the most exciting research in medicine and biology. Many believe this combination will dominate research in the next decade. If so, our difficulty in crafting legislation and social policies to take fullest advantage of this amazing opportunity are striking. This makes me believe that development of wise public policy for science should get some attention in even the sacred corridors of our medical schools.

REFERENCES

1. Lindberg DAB. The computer and medical care. Springfield: Thomas, 1968;210.
2. Lindberg DAB. The growth of medical information systems in the United States. Lexington: D.C. Heath and Company, Lexington Books, 1979;194.
3. McCray AT. The nature of lexical knowledge. Methods Inform Med. 1998 Nov;37(4–5):353–60.
4. The National Research and Education Network Program: A Report to Congress. December 1992. Washington, D.C.: Office of Science and Technology Policy.
5. High Performance Computing and Communications: Foundation for America's Information Future. Supplement to the President's 1996 Budget. Washington, D.C. National Science and Technology Council.
6. Public Law 100-607, November 4, 1988.
7. Spitzer V, Ackerman MJ, Scherzinger AL, Whitlock D. The Visible Human Male: A Technical Report. JAMIA, 1996;3(2):118–130.
8. Ackerman, MJ. The Visible Human Project. Proc. I.E.E.E. 1998;86(3):504–511.
9. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inform Med. 1993 Aug;32(4):281–291.
10. Selden CR, Humphreys BL. Unified Medical Language System. Current Bibliographies in Medicine. 1997;8:96.
11. Cooper JM, Katcher MS, Orlov MV. Implantable devices for the treatment of atrial fibrillation. N Engl J Med 2002 Jun 27;346(26):2062–2068.
12. Williams & Wilkins Co. v. United States, 487 F.2d 1345 (Ct. Cl. 1973), affirmed by an equally divided Court, 420 U.S. 376 (1975).
13. Lacroix E-M, Mehnert, R. The US National Library of Medicine in the 21st Century: Expanding Collections, Nontraditional Formats, New audiences. Health Information and Libraries Journal. 2002 Sep, 19(3):126–132.
14. Sultson J, Ferry G. The Common Thread. Washington, D.C.: Joseph Henry Press; 2002;144–147.
15. For the Record: Protecting Electronic Health Information. Washington, D.C.: National Academy Press 1997;264 pp.
16. Medical Privacy Stories. Washington, D.C.: Institute for Health Care Research and Policy, Georgetown University. 2002. www.healthprivacy.org/content2310/content.htm.
17. Sweeney L. Foundations of Privacy Protection from a Computer Science Perspective. Proceedings, Joint Statistical Meeting, American Association for the Advancement of Science, Indianapolis, IN. 2000.
18. Bioterrorism on the home front. A new challenge for American Medicine. HC Lane, AS Fauci; JAMA; v 286 #20, Nov 28, 2001.

DISCUSSION

Jameson, Chicago: Since the web engine "Google" was discussed earlier this morning, are there any strategies that you can think of, or that you use to try to target the higher quality or more reliable information to such search engines to provide users with information from responsible professional organizations versus information from anyone who wants to post information on the web?

Lindberg: Bethesda: I take the point, how do you help people distinguish between the two. I say, go to the NLM, the National Library of Medicine first (PubMed for doctors, MedlinePlus for patients, families, and the public), because that's going to be 100% reliable, and will cite many other reliable sources. I actually did a Google search for GATA 1. Google also listed NCBI at NLM as the top reference for GATA 1. I'm not sure how they did that. It's not a matter of Google being better or worse than other search engines. We are in the sense of Internet back in the times when books appeared. There was and is lots of nonsense published in books and magazines, and no one seems to worry about it as much as they do with the Internet. I think you have to get a little bit sophisticated and pay attention to who's responsible for these Internet sites, how often are they updated? Who's paying for them? Are they selling things to you? Do they demand your identity?

Jameson: I guess my question is; is there anything you can do at the National Library of Medicine that will prioritize your information for sites like Google, given that you understand how Google and similar sites are structured.

Lindberg: In general, people buy their way into high priority on these commercial search engines. We did not. I wouldn't be legally permitted to, but I wouldn't do it if I could. I don't know how Google figured out that the best place to come for GATA 1 information is NCBI at NLM, I really don't, but probably by our word placement and word frequency.