

# Summary of Research Activities by Key Approach and Resource

## Disease Registries, Databases, and Biomedical Information Systems

*The Atlanta physician had never before treated a boy complaining of “numb chin.” He sent the lad to the examining room to undress and quickly turned to his personal computer. He typed in the term “numb chin” and read from the screen a lengthy description of an article on just that subject. This information provided the vital clue the physician needed to diagnose a form of lymphoma. Similar scenes are reenacted thousands of times every day in physicians’ offices, research laboratories, hospital nurses’ stations, and medical schools—in short, wherever health professionals require information. The system this physician tapped into is MEDLINE, just one of the National Library of Medicine’s growing numbers of online databases containing medical references, abstracts, and other essential health information for professionals as well as the general public.*

### Introduction

From bench to bedside and from database to desktop, information itself has become a primary driver of progress in the biomedical and health care enterprise. For example, the volumes of data resulting from sequencing the genomes of thousands of patients have become primary resources for identifying associations between specific genes and diseases. The data that flow from large-scale clinical studies, advanced diagnostic and imaging equipment, and electronic medical records is a key enabler of improvements in clinical practice and individual patient care. Up-to-date information from disease registries has become a critical resource for studying disease incidence and treatment patterns and forms the basis for public health interventions. The availability of this and other health information on the Internet enables consumers to play a more active role in managing their health and further increases demand for reliable and authoritative health information.

The development, deployment, and utilization of disease registries, databases, and other biomedical information systems are essential to managing large amounts of data for improved health. Such systems permit the efficient collection, storage, and accessing of biomedical information. Disease registries collect information about the occurrence of specific diseases, such as cancer and Parkinson's disease, the kinds of treatment that registered patients receive, and other information that might be relevant to researchers or public health officials. This information can help in identifying causal factors of disease, assessing the effectiveness of various interventions, and identifying questions of concern to researchers, clinical professionals, and policymakers. Biomedical databases serve as repositories for a wide range of information, from the results of scientific or clinical research studies, to genomic information, to standard reference materials (such as genome sequences or anatomical images), to published journal articles and citations to the medical literature. They are widely used by biomedical researchers, as well as a growing number of clinicians, public health officials, and consumers.

Increasingly, biomedical databases serve not only as repositories of information, but also as research tools in and of themselves. Discoveries can be made by examining the information they contain. Scientists can use molecular databases to study the molecular profiles of individual tumors and create small-molecule anticancer agents to target them. They can analyze large-scale databases linking genotype and phenotype information from thousands of individuals to identify the genes associated with particular observable traits (e.g., obesity) or diseases (e.g., diabetes, cancer). In these ways, biomedical information systems are changing the nature of research itself, and promise to change the nature of clinical care and public health.

The utility of biomedical information systems rests on many factors, including the quality of the data they contain,

accessibility to the full range of potential users, how easily they can be searched to find relevant and interesting results, and the availability of useful tools for analyzing the information they contain. New data must be added on a regular basis, while existing data are maintained or updated to reflect new findings. Improved search tools are needed to comb through the massive datasets and retrieve relevant results. Standard vocabularies that are used to organize information and ensure accurate retrieval must be updated to accommodate new concepts and relationships. New analytical tools are needed to explore increasingly complex questions, such as how the expression patterns of multiple genes are associated with a particular trait or response. Preserving, protecting, and ensuring the validity and security of information stored in biomedical databases remains of paramount importance.

## Scope of NIH Activities in Disease Registries, Databases, and Biomedical Information Systems

Because of the growing importance of information and its management in biomedical science, clinical care, and public health, virtually every NIH IC is engaged in the development, deployment, and use of biomedical information systems that support their mission. Several trans-NIH Roadmap activities also feature the development of significant biomedical information resources, including the tools, infrastructure, and associated research needed to make databases and registries more valuable.

Most evident among NIH's biomedical information resources are major scientific databases such as [GenBank](#) (genomic sequence data), [PubChem](#) (small molecules data), and [dbGaP](#) (database of Genotype and Phenotype). These and many similar databases (NLM alone oversees 36) have become indispensable national and international resources for biomedical and health research and public health. DNA sequence data stored at NIH, for example, allowed rapid identification of the first known polio case in the United States since 1999 and the rapid initiation of treatment.

NIH also houses the leading source of authoritative biomedical literature for professional and lay audiences. NLM's exhaustive [Medline/PubMed](#) database, for example, indexes citations to some 5,000 peer-reviewed scientific journals on a regular basis. [PubMed Central](#), its digital archive of full-text articles, provides online access to a growing number of scientific journal articles deposited by publishers and by NIH-funded researchers who are complying with the [NIH Public Access Policy](#). These comprehensive resources are widely used by scientists, health care providers, and consumers who seek peer-reviewed information on biomedical and health topics of interest. As another example, peer-reviewed studies from the [National Toxicology Program](#) are used by State, local, and Federal health officials to assess the toxicologic potential of environmental compounds to cause adverse health effects such as cancer.

NIH also works with other Federal and private entities to integrate disease registries for national and local use. For example, for 34 years the [Surveillance, Epidemiology, and End Results](#) (SEER) program has collected and published cancer incidence and survival data from cancer registries covering approximately 26 percent of the American population. SEER information has been the foundation for innumerable studies, including recent research into links between hormone therapy and breast cancer. [NIAMS](#) supports a dozen [registries](#) associated with specific diseases, including lupus, muscular dystrophy, and rheumatoid arthritis.

NIH's work on biomedical information systems goes beyond the establishment of databases and registries. NIH is also the largest Federal funder of biomedical informatics research, which aims to advance the applications of computing to biomedicine—for both research and clinical care. Grant programs support research and training in medical informatics and medical librarianship. NIH also leads the government's efforts to develop standardized vocabularies and terminology to support interoperability among biomedical information systems and has developed numerous tools to facilitate data analysis. These efforts aim to create and sustain the biomedical information infrastructure needed for research, clinical care (including electronic health records), and public health.

# Summary of NIH Activities

## Expanding and Enhancing NIH Scientific Databases

Keeping pace with the expanding biomedical knowledge base is a continuing challenge for scientists; thus, NIH devotes considerable attention and resources to developing, expanding, and maintaining tools and resources for information management. NLM's Medline/PubMed database of the peer-reviewed bioscience literature, for example, added almost 1.3 million new citations to its archive in the 2-year period of FYs 2006-2007 and now contains indexed citations of more than 17 million articles, editorials, comments, and other materials. PubMed Central, NIH's electronic archive of full-text journal articles, passed the 1 million article mark in June 2007. NIH's online registry of clinical trials, [Clinicaltrials.gov](http://Clinicaltrials.gov), added information on more than 23,000 clinical research studies in FY 2006-2007 and by the end of FY 2007 contained information on some 47,000 clinical trials conducted in more than 140 countries. Many other NIH databases have seen similar growth, placing greater demand on NIH's information infrastructure and on the resources needed to input, store, and index information. Ongoing efforts are needed to streamline such processes and boost their productivity.

Increased utilization goes hand in hand with the expanding content of NIH's databases. [Medline](http://Medline) alone logged nearly 900 million searches in FY 2006, almost twice the level of FY 2003, and [Clinicaltrials.gov](http://Clinicaltrials.gov) saw some 500 thousand unique visitors in June 2007, double the number 2 years earlier. Some of this increase is attributable to an expanding scope of users—not just biomedical researchers, but also clinicians, consumers, and other practitioners. NIH actively endeavors to make its information resources more accessible to varied types of users, as illustrated by its work on [MedlinePlus](http://MedlinePlus), NLM's comprehensive health information source for consumers and health professionals, and [WISER](http://WISER), the Wireless Information System for Emergency Responders. WISER makes information available to emergency responders from NIH's [TOXNET](http://TOXNET) databases. TOXNET, is a cluster of 14 large databases covering toxicology, hazardous chemicals, environmental health, and related topics. It has been used by toxicologists for decades, assisting them in locating toxicology data, literature references, and toxic release information on particular chemicals, as well as in identifying chemicals that cause specific health effects. To make these resources more useful to first responders at the scene of a disaster, NIH developed the WISER system, which enables wireless access to a selection of the most relevant data for emergency responders. WISER can be installed on personal digital assistants, providing emergency personnel with access to critical information for identifying and safely cleaning up spilled chemicals, understanding their health effects, treating exposed victims, and assessing environmental impact. A Web-based version of WISER, released in FY 2006, can be accessed from any Web-enabled device. WISER was used to access information on dangerous chemicals to which Hurricane Katrina victims may have been exposed.

## Genomic Information Systems

NIH also has made great strides in developing information resources to support genetics research. NIH has long supported genetic information resources through widely used repositories such as [GenBank](http://GenBank), the NIH genetic sequence database. More recent efforts have aimed to support the analysis of data from genome-wide association studies, which explore the connection between specific genes (genotype information) and phenotype information, such as observable traits (e.g., blood pressure, weight) or a particular disease. NIH's [dbGaP](http://dbGaP) (database of Genotype and Phenotype) was launched in December 2006 to house data from a number of genome-wide association studies. By the end of FY 2007, dbGaP contained more than 12 large datasets, including genetic analyses from the landmark [Framingham Heart Study](http://Framingham Heart Study), and studies of age-related eye diseases and Parkinson's disease. The wide availability of information linking genotype to phenotype should help researchers better understand gene-based diseases and speed development of effective therapies.

Several NIH ICs have established genetics repositories to accelerate research and multidisciplinary collaborations in specific disease areas. Programs such as the [NIMH Genetics Repository](http://NIMH Genetics Repository), the NINDS [Human Genetics Repository](http://Human Genetics Repository), the NCBI [Influenza Virus Resource](http://Influenza Virus Resource), the NIA [Genetics of Alzheimer's Disease Data Storage Site](http://Genetics of Alzheimer's Disease Data Storage Site), and the [National Database for Autism Research](http://National Database for Autism Research) give researchers access to vast storehouses of genetic and genomic data, DNA

samples, and clinical data, along with informatics tools designed to facilitate their analyses. For example, the Influenza Virus Resource database, comprising information obtained from the NIAID Influenza Genome Sequencing Project and GenBank, contains more than 40,000 influenza virus sequences, including the sequences of more than 2,500 whole influenza genomes. More than 11,000 sequences were added in FY 2006, along with new search and annotation tools. This resource enables scientists to compare influenza virus strains so that emergent variants can be more rapidly identified and vaccines developed accordingly. As the library of viral sequences grows it will be an increasingly important reference to help further understand how avian viruses spread to humans, and how influenza activity spreads throughout the world.

## Disease Registries and Surveillance Systems

NIH-supported disease registries have paid many dividends over the years. Recently, for example, with the participation of patients from the [Alopecia Areata Registry](#), NIH-supported scientists discovered four chromosomal locations that appear to be associated with susceptibility to this common autoimmune disease, which is characterized by patchy hair loss. Understanding the mechanisms of the genes found at these locations could lead to the development of an effective treatment for the disease, which is presently untreatable.

Registries also serve as an effective mechanism to gather data on the incidence, prevalence, and natural history of diseases. The NIEHS-supported [California Parkinson's Disease Registry](#), for example, enables researchers to identify the possible environmental and genetic origins of this progressive neurological disorder suffered by an estimated 1.5 million Americans. Data in the registry can help to determine whether race, ethnicity, gender, age, environmental factors, or place of residence influence the likelihood of getting the disease, and can help track incidence and demographic trends.

Registries also are integral elements of more comprehensive NIH programs designed to monitor and analyze disease trends in the United States. For example, the [Surveillance, Epidemiology, and End Results \(SEER\)](#) program has a rich track record of identifying emerging trends, geographic variation, ethnic disparities, and other patterns that have provided new directions for epidemiologic research in cancer etiology and control. SEER data provided critical insight into the relationship between hormone therapy and breast cancer incident rates. Reported incidents of breast cancer in the SEER registry began to decline in mid-2002, shortly after a highly publicized series of reports from the NIH [Women's Health Initiative](#) (WHI) revealed an association between the risk of breast cancer and the use of hormone therapy. By analyzing SEER data on breast cancer incidence rates using several key factors such as the estrogen-receptor status of tumors, WHI researchers demonstrated that the incidence of tumors most likely to be affected by changes in hormone therapy reflected usage patterns, while trends for other tumors did not.

Surveillance and monitoring programs are also crucial sources of information and analysis for policymakers, legislators, public health officials, clinicians, and the public. SEER participates in [Cancer Control P.L.A.N.E.T.](#) (Plan, Link, Act, Network with Evidence-based Tools), a Web portal that provides links to comprehensive cancer control resources and data for public health professionals. NIDA supports several epidemiologic programs designed to gather ongoing data and monitor emerging drug abuse trends in adolescents and other populations, helping to guide national and global prevention efforts, drug control, and public health policy. Among the projects is the [Monitoring the Future \(MTF\) Survey](#), which has been tracking trends in substance use, attitudes, and beliefs among adolescents and young adults in the United States since 1975. Data from the 2007 MTF Survey show good news and continuing areas of concern. For although teen drug use continues to decline—including cigarette smoking, now at the lowest rate in the survey's history—use of prescription-type drugs is still high, with more than 15 percent of 12th graders reporting nonmedical use within the past year.

## Enhancing the Utility of Data Resources: Tools and Standards

Other efforts aim to enhance the utility of NIH databases. A key element of this work is to exploit the inherent relationships among information in disparate databases. NIH's [PubChem](#) database, for example, is an integrated hub within the Entrez suite of biomedical information resources. PubChem is the repository for data flowing from the high-throughput bioassay centers that were established with NIH funding under the [Molecular Libraries Initiative](#) of the NIH Roadmap. It provides information about the biological activity of small molecules, organized as

three linked databases along with a chemical structure similarity search tool. PubChem's chemical structure and bioassay records are interlinked with the biomedical literature in PubMed and with three-dimensional protein structure records. This integration provides many routes by which biomedical researchers may discover the candidate probes developed by the Molecular Libraries Initiative. A researcher examining a protein sequence record, for example, may see that a particular protein has been screened, view the active compounds, and examine structure-activity relationships using PubChem analysis tools. NLM's Discovery Initiative, launched in FY 2006-2007, aims to take database linking to the next level. The Discovery Initiative will improve the presentation of results from search queries conducted across a range of NIH databases so that users, who often do not go beyond retrieving the basic results of a search query, are more likely to be drawn to related information that could lead to serendipitous discoveries, even if that information resides in another NIH database.

Other efforts relate to the development of standardized nomenclatures and data protocols. Medical terminology can be difficult to remember and can vary from one laboratory or clinical facility to another. Often there are many names for a single concept (e.g., cancer of the colon, colonic neoplasm, colon cancer). Standard vocabularies and ontologies (models of the relationships between concepts) improve information search and retrieval by endowing systems with the ability to automatically perceive and retrieve information about related terms. As expansion of the scientific frontiers produces new concepts, terms, and relationships, standard vocabularies must be regularly revised so that articles and other data can be properly indexed and search engines can find relevant and related terms.

NLM continues to update its Unified Medical Language System (UMLS), which is heavily used in advanced biomedical research and data mining worldwide. NLM and many other institutions apply UMLS resources in a wide variety of [applications](#) including information retrieval, natural language processing, creation of patient and research data, and the development of enterprise-wide vocabulary services. NIH's [ClinicalTrials.gov](#) database now uses the UMLS to improve the system's ability to retrieve information about clinical trials related to a user's interests.

UMLS and related NIH programs also contribute to efforts to national efforts to expand the use of electronic health records to improve the quality and efficiency of health care. Standardized clinical terminology and coding systems facilitate the exchange of information among care providers, insurers, and patients, contributing to implementation of an interoperable health information technology infrastructure. NLM is the government's lead agency for maintaining and disseminating clinical terminology standards. In 2007, NIH helped to establish the International Health Terminology Standards Development Organization, which is globally distributing SNOMED CT (Systematized Nomenclature of Medicine?Clinical Terms), a comprehensive clinical terminology for electronic health records.

## **Informatics/Computational Biology Initiatives**

NIH also has embarked on a number of large-scale initiatives to develop and deploy infrastructure and tools for storing, sharing, integrating, and analyzing the large volumes of data routinely generated by today's laboratories.

In the area of cancer research, for example, NIH has established the cancer Biomedical Informatics Grid ([caBIG](#)). caBIG is a collaborative information network for all of NCI's advanced technology and program initiatives, connecting scientists, practitioners, and patients and enabling the collection, analysis, and sharing of data and knowledge along the entire research pathway from bench to bedside. Specific biomedical research tools under development by caBIG include clinical trial management systems, tissue repositories and pathology tools, imaging tools, and a rich collection of integrative cancer research applications. Patients benefit from caBIG through systems and services such as [BreastCancerTrials.org](#) (BCT), which was launched in 2006 to match patients' medical case histories to ongoing clinical trials in the greater San Francisco and Sacramento areas. Created by patients for patients, BCT is an online version of a caBIG tool called [caMatch](#), which aims to save patients time and energy, while also giving them greater options in seeking clinical trials that are relevant to their condition.

Other efforts aim to provide the informatics infrastructure to advance basic research and clinical studies across the

spectrum of biomedical sciences. NIH's [Biomedical Informatics Research Network](#), for example, is a virtual community of shared informatics resources. It includes a data repository that makes research data freely available for sharing and exchange; data integration tools that allow searching across distributed databases; and tools for data analysis, management, and collaborative research. The [National Electronic Clinical Trials and Research \(NECTAR\)](#) network, a clinical research “network of networks,” is a Roadmap initiative that will provide the informatics infrastructure for interconnected and inter-operable clinical research networks. NECTAR will allow clinical investigators to broaden the scope of their research while enhancing efficiency and reducing duplication of efforts by integrating clinical research networks that currently operate independently of each other. Another Roadmap initiative will create a national software engineering system through Bioinformatics and Computational Biology initiatives. Through a computer-based grid, biologists, chemists, physicists, computer scientists, and physicians anywhere in the country will be able to share and analyze data using a common set of software tools. The National Centers for Biomedical Computing are a central focus of this effort.

## Biomedical Informatics Research and Training

Ensuring continued advances in biomedical informatics resources requires active support of fundamental research that seeds the further development of new tools, resources, and approaches. It also is critical to generating a continuous supply of skilled biomedical informatics researchers, information specialists (such as medical librarians), and life sciences researchers trained in bioinformatics. NIH continues to expand its efforts in bioinformatics research and training in response to the growing importance of informatics in the biomedical and life sciences (see section on Research Training).

Several ICs fund informatics research projects within their areas of specialization. However, NLM remains the primary Federal sponsor of biomedical informatics research, and its extramural grants program supports research on the characterization, management, and efficient use of data, information, and knowledge in health care and basic biomedical sciences. Grants funded in FY 2006-2007 explored informatics challenges related to clinical care, biomedical research, genomics, and public health. NLM's long-range plan, *Charting a Course for the 21st Century*, published in September 2006, identifies a number of emerging informatics challenges that will demand continued research and development.

## Notable Examples of NIH Activity

### Key for Bulleted Items:

E = Supported through Extramural research

I = Supported through Intramural research

O = Other (e.g., policy, planning, and communication)

COE = Supported through a congressionally mandated Center of Excellence program

GPRA Goal = Concerns progress tracked under the Government Performance and Results Act

## Scientific Databases

**MEDLINE/PubMed and PubMed Central (PMC):** NIH continued to expand MEDLINE/PubMed as a tool for biomedical research, clinical medicine and consumer health. Almost 1.3 million citations were added to MEDLINE/PubMed in FY 2006-2007, a 10 percent increase from the previous two-year period. NIH made significant strides in enhancing PMC, its repository of full-text biomedical journal articles. PMC surpassed the 1 million-articles mark in June 2007, and, to support NIH policy on public access to NIH-funded research, the NIH Manuscript Submission system was developed, enabling NIH grantees to deposit manuscripts into PMC. To foster international cooperation on preservation and access to biomedical literature, NIH made PMC software available to archiving

organizations outside the United States and worked with the Wellcome Trust and other major United Kingdom research funders in the to establish a UKPMC service. Five other countries plan to establish PMC sites.

- For more information, see <http://www.pubmed.gov>
- For more information, see <http://www.nihms.nih.gov/>
- For more information, see <http://www.pubmedcentral.nih.gov/about/pmci.html>
- For more information, see <http://ukpmc.ac.uk/>
- (I) (NLM)

**MedlinePlus/MedlinePlus en Español:** NIH employed new methods to increase awareness of its MedlinePlus databases. Weekly podcasts by NLM's Director were initiated to provide timely reports on health news; *NIH MedlinePlus The Magazine* was rolled out at a press event on Capitol Hill attended by members of Congress and guest celebrity Mary Tyler Moore, featured on the cover. The magazine is distributed free of charge to 40,000 physician offices and has covered stories on cancer, diabetes, and heart attack. NIH expanded the content and features of the English and Spanish MedlinePlus Web sites and the associated GoLocal sites that provide information on local health resources for approximately one-third of the U.S. population. MedlinePlus was one of two U.S. winners of the 2005 Award at the World Summit on the Information Society.

- For more information, see <http://www.medlineplus.gov>
- This example also appears in Chapter 3: *Health Communication and Information Campaigns and Clearinghouses*.
- (I) (NLM)

**Toxicology Data NETWORK (TOXNET):** TOXNET is a cluster of more than 10 databases covering toxicology, hazardous chemicals, environmental health, and related topics. It is a primary reference for toxicologists, poison control centers, public health administrators, physicians, and other environmental health professionals. In 2006, the Hazardous Substances Data Bank, which contains comprehensive information on more than 5,000 substances, was expanded to include a general record for [ionizing radiation](#) and a series of specific radionuclide records. In 2007, LactMed, a peer-reviewed and fully referenced database of drugs to which breast-feeding mothers may be exposed, was added to TOXNET.

- [Wexler P. \*Toxicology\* 2004;198:161-8](#), PMID: 15138039
- [Tomasulo P. \*Med Ref Serv Q\* 2007 Spring;26:51-8](#), PMID: 17210549
- For more information, see <http://toxnet.nlm.nih.gov>
- (I) (NLM)

**National NeuroAids Tissue Consortium (NNTC):** The NNTC is a repository of brain tissue and fluids from highly characterized HIV+ individuals. Established as a resource for the research community, the NNTC includes information from over 2,000 individuals, including approximately 641 brains, thousands of plasma and cerebrospinal fluid samples, and additional organs and nerves of interest.

- For more information, see <http://grants1.nih.gov/grants/guide/rfa-files/RFA-MH-08-021.html>
- This example also appears in Chapter 2: *Neuroscience and Disorders of the Nervous System*
- (E/I) (NIMH, NINDS)

**ClinicalTrials.gov:** Established in 2000 in response to congressional mandate (Pub. L. No. 105-115), ClinicalTrials.gov has grown to become the largest clinical trial registry in the world with information on clinical research studies for hundreds of diseases and conditions conducted in 148 countries. At the end of September 2007, it contained more than 47,000 registered trials—more than double the number of entries 2 years earlier. Legislation enacted in September 2007, the Food and Drug Administration Amendments Act of 2007 (Pub. L. No. 110-85), expanded the scope of trials to be registered with ClinicalTrials.gov and the registration information to be provided. It also mandates the inclusion of specified results information beginning in September 2008.

- [Drazen JM, et al. \*N Engl J Med\* 2007;356:184-5](#), PMID: 17215537
- [Zarin DA, et al. \*N Engl J Med\* 2005;353:2779-87](#), PMID: 16382064
- For more information, see <http://clinicaltrials.gov>
- This example also appears in Chapter 3: *Clinical and Translational Research*.
- (I) (NLM)

**Influenza Virus Resource:** This database of more than 40,000 influenza virus sequences allows researchers around the world to compare different virus strains, identify genetic factors that determine the virulence of virus strains, and look for new therapeutic, diagnostic, and vaccine targets. The resource was developed by NCBI using data obtained from NCBI's Influenza Virus Sequence Database and from NIAID's Influenza Genome Sequencing Project, which has contributed sequences of the complete genomes from over 2,500 influenza samples. In FY 2006 more than 11,000 influenza virus sequences were entered into the database, and new search and annotation tools were added to assist researchers in their analyses.

- [Wolf YI, et al. \*Biol Direct\* 2006;1:34](#), PMID: 17067369
- [Chang S, et al. \*Nucleic Acids Res\* 2007;35:D376-80](#), PMID: 17065465
- For more information, see <http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html>
- For more information, see <http://www.niaid.nih.gov/dmid/genomes/mscs/influenza.htm>
- This example also appears in Chapter 2: *Infectious Diseases and Biodefense*, Chapter 3: *Molecular Biology and Basic Sciences*, and Chapter 3: *Genomics*
- (1) (NLM, NIAID)

**PubChem:** PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative. By the end of 2007, PubChem contained information on more than 38 million substances, 18 million compounds, and 710 bioassays.

- For more information, see <http://pubchem.ncbi.nlm.nih.gov/>
- (E) (Roadmap—all ICs participate)

**Databases for Cervical Cancer Research:** NIH has developed data analysis and image recognition tools for studying biomedical images of human papillomavirus (HPV) infection and cervical neoplasia. Image data include 100,000 cervicographs (high-definition cervical photograph), Pap test, and histology images. Tools allow the exploration of visual aspects of HPV and cervical cancer for research, training, and teaching.

- [Castle PE, et al. \*Cancer Res\* 2006;66:1218-24](#), PMID: 16424061
- [Jeronimo J, et al. \*J Low Genit Tract Dis\* 2006;10:39-44](#), PMID: 16378030
- This example also appears in Chapter 3: *Epidemiological and Longitudinal Studies* and Chapter 2: *Cancer*.
- (I) (NLM, NCI)

## Genomic Information Systems

**Database of Genotype and Phenotype (dbGaP):** Research on the connection between genetics and human health and disease has grown exponentially since completion of the Human Genome Project in 2003, generating high volumes of data. Building on its established research resources in genetics, genomics and other scientific data, NIH established dbGaP to house this growing body of information, particularly the results of GWAS, which examine genetic data of subjects with and without a disease or specific trait to identify potentially causative genes. By the end of 2007, dbGaP included results from more than a dozen GWAS, including genetic analyses added to the landmark Framingham Heart Study and trials conducted under the Genetic Association Information Network. dbGaP is to become the central repository for many NIH-funded GWAS in order to provide for rapid and widespread distribution of such data to researchers and accelerate the advance of personalized medicine.

- For more information, see <http://view.ncbi.nlm.nih.gov/dbgap>
- This example also appears in Chapter 3: *Epidemiological and Longitudinal Studies* and Chapter 3: *Genomics*.
- (I) (NLM)

**Genome-Wide Association Studies (GWAS) and Database of Genotype and Phenotype (dbGaP):** In December 2006, NIH released the initial dbGaP dataset using genome-wide association data from the Age-Related Eye Diseases Study (AREDS), a landmark study of the clinical course of age-related macular degeneration (AMD) and cataracts. AREDS documents, protocols, and aggregated data are made available with no restrictions. In order to protect patient confidentiality, de-identified individual-level patient characteristics and family data are accessible only by authorized investigators. Correlating phenotype and genotype data provides information about the genetic and environmental interactions involved in a disease process or condition, which is critical for better understanding complex diseases and developing new diagnostic methods and treatments. Using these data, recent studies have linked two genes with progression to advanced AMD. After controlling for other factors, certain forms of the genes increased risk of AMD progression 2.6- to 4.1-fold; smoking and body weight further increased risk with these gene variants.

- [Seddon JM, et al. JAMA 2007;297:1793-800](#), PMID: 17456821
- For more information, see <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap>
- This example also appears in Chapter 2: *Chronic Diseases and Organ Systems* and Chapter 3: *Genomics*.
- (E) (NEI, NIA, NLM)

**NIMH Genetics Repository:** Over the last 9 years, NIMH has built the infrastructure for large-scale genetics studies through the NIMH Human Genetics Initiative. Through this Initiative, NIMH established a repository of DNA, cell cultures, and clinical data-serving as a national resource for researchers studying the genetics of complex mental disorders.

- For more information, see <http://nimhgenetics.org/>
- This example also appears in Chapter 3: *Genomics* and Chapter 2: *Neuroscience and Disorders of the Nervous System*.
- (E) (NIMH)

**NINDS Human Genetics Repository:** In 2003, NINDS established this Repository to collect, store, characterize, and distribute DNA samples and cell lines and standardized clinical data for the research community. By June 2007, the repository held material from 16,683 subjects, including stroke (4,363), epilepsy (1,065), Parkinson's disease (3,585), motor neuron diseases such as ALS (2,445), and control samples (4,767). The ethnically diverse collection represents populations from the United States and several other countries. Investigators have submitted or published more than 50 scientific articles based on data from this resource, and technological advances allowing "whole genome screening" for disease genes have also enhanced its value.

- For more information, see <http://ccr.coriell.org/Sections/Collections/NINDS?Sslid=10>
- This example also appears in Chapter 2: *Neuroscience and Disorders of the Nervous System*.
- (E/I) (NINDS)

**Candidate Gene-Association Resource:** Over the years, NHLBI has supported a number of major population studies that have collected extensive data on cardiovascular disease and its risk factors and manifestations. To increase the utility of the data for conducting genetic association studies, NIH initiated the Candidate Gene Association Resource program in FY 2006. This new resource will have the capacity to perform high-throughput genotyping for up to 50,000 subjects in cohort studies that have stored samples and data available on a wide array of characteristics (phenotypes) associated with heart, lung, blood, and sleep disorders. The linked genotype-phenotype data will form an invaluable resource for investigators seeking to identify genetic variants related to those disorders.

- For more information, see <http://public.nhlbi.nih.gov/GeneticsGenomics/home/care.aspx>
- This example also appears in Chapter 2: Chronic Diseases and Organ Systems and Chapter 3: Genomics.
- (E) (NHLBI)

**Alzheimer's Disease (AD) Genetics Initiative and Data Storage:** Only one of the four validated AD genes, APOE, has been definitively linked with the more common late-onset form of the disease. A fifth gene, SORL1, has recently been linked with late-onset Alzheimer's disease (LOAD) in some studies. The goal of the AD Genetics Initiative is to develop the resources necessary for identifying the LOAD risk factor genes and the interactions of genes with the environment. In FY 2006, NIH achieved its goal to recruit 1,000 families with two or more siblings living with AD through an unprecedented alliance of AD Centers, researchers, and outreach with the Alzheimer's Association. To facilitate access by qualified investigators, all genetic data derived from NIH-funded studies on LOAD genetics are deposited at a central data storage site at Washington University in St. Louis, another NIH-approved site, or both. Discovery of risk factor genes will help illuminate the underlying disease processes of AD, open up novel areas of research, and identify new targets for drug therapy.

- For more information, see <http://www.niageneticsdata.org/>
- This example also appears in Chapter 2: Neuroscience and Disorders of the Nervous System.
- (E/I) (NIA)

**Autoimmune Diseases and Genetics:** With the advancement of genomic science, more information has been gained about the genetic component of autoimmune diseases. Susceptibility genes have been identified for rheumatoid arthritis, lupus, psoriasis, and alopecia areata. Understanding the genetic influence of these diseases provides essential information for the design of new therapies.

- [Kumar KR, et al. Science 2006;312:1665-9](#), PMID: 16778059
- [Nair RP, et al. Am J Hum Genet 2006;78:827-51](#), PMID: 16642438
- [Haas CS, et al. Arthritis Rheum 2006;54:2047-60](#), PMID: 16804865
- [Martinez-Mir A, et al. Am J Hum Genet 2007;80:316-28](#), PMID: 17236136
- For more information, see [http://www.niams.nih.gov/News\\_and\\_Events/Spotlight\\_on\\_Research/2006/lupus\\_susceptibility\\_gene.asp](http://www.niams.nih.gov/News_and_Events/Spotlight_on_Research/2006/lupus_susceptibility_gene.asp)
- This example also appears in Chapter 2: *Autoimmune Diseases*.
- (E) (NIAMS, NCR, NHLBI, NIAID, NIMH)

## Disease Registries and Surveillance Systems

**Interagency Registry for Mechanically Assisted Circulatory Support (INTERMACS):** In a joint effort, NHLBI, the Center for Medicare & Medicaid Services, and the U.S. Food and Drug Administration created INTERMACS, a national registry for patients who are receiving mechanical circulatory support device therapy to treat advanced heart failure. Data from INTERMACS are expected to improve patient evaluation and management, aid in the development of safer, more effective devices, and enhance research.

- For more information, see <http://www.uab.edu/ctsresearch/mcsd/>
- This example also appears in Chapter 2: Chronic Diseases and Organ Systems and Chapter 3: Technology Development.
- (E) (NHLBI)

**Resuscitation Outcomes Consortium:** Recognizing the critical importance of early intervention for victims of cardiopulmonary arrest and traumatic injury, in FY 2004 NIH and its U.S. and Canadian partners initiated the Resuscitation Outcomes Consortium, a large-scale network to conduct clinical trials of promising approaches to improving outcomes. During FY 2006-2007, two Consortium clinical trials began enrolling patients—one to compare the efficacy of three fluids for initial resuscitation of hypotensive or brain-injured patients and the other to test two strategies for increasing blood flow during cardiopulmonary resuscitation. The Consortium also

established a pre-hospital Cardiac Arrest and Trauma Registry across the United States and Canada. In addition, emergency medicine fellowship training programs established at several study sites are enhancing training in resuscitation medicine.

- For more information, see <https://roc.uwctc.org/>
- This example also appears in Chapter 3: Clinical and Translational Research.
- (E) (NHLBI, NINDS)

**A Look at Drug Abuse Trends: Local to International:** Several major systems of data collection are helping to identify substance abuse trends locally, nationally, and internationally: Monitoring the Future Survey (MTF), the Community Epidemiology Work Group (CEWG), and the Border Epidemiology Work Group (BEWG). All help to surface emerging drug abuse trends among adolescents and other populations, and guide responsive national and global prevention efforts. The MTF project, begun in 1975, has many purposes, the primary one being to track trends in substance use, attitudes, and beliefs among adolescents and young adults. The survey findings are also used by the President's Office of National Drug Control Policy to monitor progress towards national health goals. The MTF project includes both cross-sectional and longitudinal formats—the former given annually to 8th, 10th, and 12th graders to see how answers change over time, and the latter given biennially, or every 2 years (until age 30, then every 5 years) to follow up on a randomly selected sample from each senior class. CEWG, established in 1976, provides both national and international information about drug abuse trends through a network of researchers from different geographic areas. Regular meetings feature presentations on selected topics, as well as those offering international perspectives on drug abuse patterns and trends. A recently established Border Epidemiology Work Group represents a collaboration of researchers from both sides of the U.S.-Mexico border. Of special interest are drug abuse patterns and problems in geographically proximal sister cities/areas. Development of a Latin American Epidemiology Network is under way. NIH has also provided technical consultation for the planning and establishment of an Asian multi-city epidemiological network on drug abuse.

- For more information, see <http://www.monitoringthefuture.org/>
- For more information, see <http://www.drugabuse.gov/about/organization/CEWG/CEWGHome.html>
- This example also appears in Chapter 3: Epidemiological and Longitudinal Studies and Chapter 2: Minority Health and Health Disparities.
- (E) (NIDA)

**Parkinson's Disease Registry:** NIEHS has begun to address the need for more precise data on the incidence and prevalence of Parkinson's disease through support of a Parkinson's disease registry in the State of California, where the large and diverse population, coupled with the wide range of exposures that exist through agriculture and other activities, provide a unique opportunity to investigate disease-environment links. The United States does not have a national health registry to supply data on Parkinson's disease, so estimates are based on sampling by individual studies in specific locales. The Parkinson's registry in California will allow us to base national estimates on a registry drawing upon a cross-section of the population in our most populous state.

- For more information, see <http://www.theipi.org/site/parkinson/section.php?id=101>
- This example also appears in Chapter 2: Neuroscience and Disorders of the Nervous System.
- (E) (NIEHS)

**Surveillance, Epidemiology, and End Results (SEER) Program and Software Analysis Tools:** The program is an authoritative source of information on cancer incidence and survival in U.S. publications, such as the Annual Report to the Nation on the Status of Cancer, or interpretation of recent declines in breast cancer incidence to inform the public, researchers, Federal and private agencies, and Congress on national cancer rates and trends. SEER is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis, patient survival, and treatment. Linkage with Medicare and other Federal databases yields information sources that are used routinely to answer major questions on quality, cost, and variability of cancer care as well as differences by racial and ethnic populations. SEER currently collects and

publishes data from approximately 26 percent of the U.S. population. The team is developing computer applications to unify cancer registration systems, to analyze and disseminate data, and to provide limited access to the public file. SEER is considered the standard for quality among cancer registries around the world.

- For more information, see <http://seer.cancer.gov>
- For more information, see <http://surveillance.cancer.gov/>
- This example also appears in Chapter 2: Cancer.
- (E) (NCI)

**Gene Expression Changes in Facioscapulohumeral Muscular Dystrophy (FSHD):** Results from a genome-wide scan of skeletal muscle biopsies suggest a link between eye blood vessel defects and muscle defects that characterize FSHD. Patient subjects were recruited from the National Registry for Myotonic Dystrophy and FSHD Patients and Family Members.

- [Osborne RJ, et al. Neurology 2007;68:569-77.](#) PMID: 17151338
- For more information, see [http://www.niams.nih.gov/Funding/Funded\\_Research/registries.asp#dystrophy](http://www.niams.nih.gov/Funding/Funded_Research/registries.asp#dystrophy)
- This example also appears in Chapter 2: Neuroscience and Disorders of the Nervous System and Chapter 3: Genomics.
- (E) (NIAMS, NCRR, NINDS)

**Genetic Susceptibility for Alopecia Areata:** Scientists supported by NIH have identified loci on four chromosomes that appear to play a role in the development of alopecia areata, an autoimmune disease characterized by hair loss that can affect the whole scalp or, in rarer cases, the entire body. Many U.S. families recruited for the study were identified through the Alopecia Areata Registry.

- [Martinez-Mir A, et al. Am J Hum Genet 2007;80:316-28,](#) PMID: 17236136
- For more information, see [http://www.niams.nih.gov/News\\_and\\_Events/Spotlight\\_on\\_Research/2007/alopecia\\_areata.asp](http://www.niams.nih.gov/News_and_Events/Spotlight_on_Research/2007/alopecia_areata.asp)
- This example also appears in Chapter 2: *Autoimmune Diseases*.
- (E) (NIAMS, NIMH)

## Enhancing the Utility of Data Resources: Tools and Standards

**A Clearinghouse for Neuroimaging Informatics Tools and Resources:** NIH understands that researchers seeking neuroimaging analysis software tools need a convenient way to find and compare useful software. Indeed, the best or most suitable neuroimaging analysis technologies for research may be hidden in someone's laboratory or some obscure corner of cyberspace. NIH is creating a Neuroimaging Informatics Tools and Resources Clearinghouse. The 14 NIH ICs that participate in the Neuroscience Blueprint have supported the development of sophisticated, high-quality neuroimaging informatics tools and resources. The clearinghouse is intended to facilitate the dissemination of those tools and resources and promote their adoption within the extended neuroimaging community. A contract has been awarded to create the clearinghouse infrastructure. The infrastructure will include a Web site that will provide not only access to tools and resources, but also ongoing opportunities for public comment in order to guide future development and enhancement of the tools. In addition to the contract award, grant awards are being made to individual extramural scientists to enable them to render their tools more suitable for this initiative. The awards will fund the enhancement of tools to make them easier to use, more broadly applicable, or more compatible with other existing tools. The clearinghouse was released to the public in October 2007.

- For more information, see <http://www.nitrc.org/>
- For more information, see <http://neuroscienceblueprint.nih.gov/>
- This example also appears in Chapter 2: *Neuroscience and Disorders of the Nervous System*

- (E) (NIBIB, NCCAM, NCRR, NEI, NIA, NIAAA, NICHD, NIDA, NIDCD, NIDCR, NIEHS, NIGMS, NIMH, NINDS, NINR, OBSSR)

**Health IT Standards:** NIH's UMLS Metathesaurus is a distribution mechanism for standard code sets and vocabularies used in health data systems. NIH supports, develops, or licenses key health terminologies to enable their free use in U.S. electronic health record systems. In 2007, NIH helped to establish the International Health Terminology Standards Development Organization to promote more cost-effective maintenance and international adoption of the SNOMED CT clinical terminology. NIH supports ongoing development and distribution of the LOINC nomenclature for laboratory tests and patient observations and produces RxNorm, a standard clinical drug vocabulary. Another NIH resource, the Daily Med, is an official distribution mechanism for FDA-approved packaging information (drug label inserts) that links to other sources of drug information, including NIH's MedlinePlus, ClinicalTrials.gov, and PubMed. More than 60,000 people subscribe to its RSS data feeds.

- For more information, see <http://www.nlm.nih.gov/healthit.html>
- (I) (NLM)

**UMLS Knowledge Sources:** NIH's Unified Medical Language System® (UMLS) aims to facilitate the development of computer systems that behave as if they understand the meaning of biomedical and health terms. The UMLS tools underpin many production information retrieval systems at NLM and elsewhere and are heavily used in advanced research in biomedical natural language processing and data-mining across the country and around the world. The most recent UMLS Metathesaurus contains more than 1.3 million biomedical concepts and 6.4 million concept names from more than 100 source vocabularies.

- For more information, see <http://www.nlm.nih.gov/research/umls/>
- (I) (NLM)

**Radiation Event Medical Management (REMM):** As a part of an effort to improve public health emergency preparedness and response, NIH and the HHS Office of the Assistant Secretary for Preparedness and Response announced in 2007 a new downloadable online diagnostic and treatment toolkit to guide health care providers during a mass casualty radiation event. The REMM toolkit includes easy-to-follow procedures for diagnosis and management of radiation contamination and exposure, guidance for the use of radiation medical countermeasures, and a variety of other features to facilitate medical responses to radiation emergencies.

- For more information, see <http://remm.nlm.gov>
- This example also appears in Chapter 2: *Infectious Diseases and Biodefense*.
- (I) (NLM)

**Patient-Reported Outcomes Measurement Information System (PROMIS):** This NIH Roadmap initiative is developing ways to measure symptoms—such as pain, fatigue, physical functioning, social role participation, and emotional distress—that influence quality of life across numerous chronic diseases.

- For more information, see <http://www.nihpromis.org/default.aspx>
- For more information, see [http://www.niams.nih.gov/News\\_and\\_Events/Announcements/2007/PROMIS\\_supp.asp](http://www.niams.nih.gov/News_and_Events/Announcements/2007/PROMIS_supp.asp)
- This example also appears in Chapter 2: *Chronic Diseases and Organ Systems*.
- (E) (Roadmap—all ICs participate)

**The Cancer Control P.L.A.N.E.T:** This Web portal is a collaboration aimed at providing access to data and resources that can help cancer control planners, health educators, program staff, and researchers design, implement, and evaluate evidence-based cancer control programs. It assists local programs with the resources that help them determine cancer risk and the cancer burden within their State. It also helps States identify potential partners and provides online resources for interpreting research findings and recommendations, and accessing products and

guidelines for planning and evaluation.

- For more information, see <http://cancercontrolplanet.cancer.gov/>
- This example also appears in Chapter 2: *Cancer*.
- (E) (NCI)

## Informatics/Computational Biology Initiatives

**Biomedical Informatics Research Network (BIRN):** Modern biomedical research generates vast amounts of diverse and complex data. Increasingly, these data are acquired in digital form, allowing sophisticated and powerful computational and informatics tools to help scientists organize, store, query, mine, analyze, view, and, in general, make better use and sense of their data. Moreover, the digital form of these data and tools makes it possible for them to be easily and widely shared across the research community at large. NIH has supported development of the BIRN infrastructure to share data and tools by federating new software tools or using the infrastructure to federate significant datasets. BIRN fosters large-scale collaborations by utilizing the capabilities of the emerging national cyberinfrastructure. The project includes a Coordinating Center at the University of California, San Diego, which serves the critical task of developing, deploying, and maintaining key infrastructure components, including high-bandwidth connectivity, grid-based security, file management and computational services, techniques to federate databases, and shared visualization and analysis environments.

- For more information, see [www.nbirn.net](http://www.nbirn.net)
- This example also appears in Chapter 3: *Technology Development*.
- (E) (NCRR)

**National Database for Autism Research (NDAR):** The NDAR is a collaborative biomedical informatics system being created by NIH to provide a national resource to support and accelerate research in autism.

- For more information, see <http://ndar.nih.gov>
- This example also appears in Chapter 2: *Life Stages, Human Development, and Rehabilitation* and Chapter 2: *Neuroscience and Disorders of the Nervous System*.
- (E/I) (NIMH, CIT, NICHD, NIDCD, NIEHS, NINDS)

**National Centers for Biomedical Computing (NCBCs):** The NIH Roadmap Bioinformatics and Computational Biology initiative provides a networked national effort to build computational tools and infrastructure for biomedical computing. The centers are devoted to all facets of biomedical computing, from basic research in computational science to providing the tools and resources that biomedical, clinical, and behavioral researchers need to do their work. The seven centers currently supported by the NIH Roadmap have made substantial progress in software development, data resources, and scientific ontologies. These advances are currently being used by the research community for studying a broad range of biological problems including cerebral palsy, autism, diabetes, asthma, Alzheimer's disease, Huntington's disease, schizophrenia, bipolar disorder, HIV/AIDS, and prostate cancer. The long-term goal of the initiative is to create a national software engineering system that will enable biomedical and clinical researchers to share and analyze data using a common set of software tools.

- For more information, see <http://nihroadmap.nih.gov/bioinformatics/>
- This example also appears in Chapter 3: *Technology Development*.
- (E) (Roadmap—all ICs participate)

## Biomedical Informatics Research and Training

**Discovery Initiative:** The Discovery Initiative aims to maximize the utility of NIH biomedical data resources by better exploiting their inter-linkages. For example, a PubChem record on a chemical structure might link to records for similar proteins, related protein structures, and relevant journal articles. Such linkages provide users with tremendous opportunities for exploration and scientific discovery but are currently underutilized. The Discovery

Initiative aims to improve the retrieval and presentation of results so that users are more readily drawn to related data that could lead to serendipitous discoveries.

- (I) (NLM)

**Informatics Training for Global Health:** Information technology is required in almost all research programs, both to access the vast information resources available internationally and to apply to research design and analysis. This program is intended to increase the capacity of developing country scientists and medical professionals to design, access, and use modern information technology in support of health sciences research. Specifically, this program supports innovative training programs for developing country biomedical and behavioral scientists and engineers, clinicians, librarians, and other health professionals to increase their capacity to access, manage, analyze, interpret, manipulate, model, display, and share biomedical information electronically. Among other skills, this will increase their ability to conduct multisite clinical trials and international disease surveillance and prevention programs.

- For more information, see [http://www.fic.nih.gov/programs/training\\_grants/itgh/index.htm](http://www.fic.nih.gov/programs/training_grants/itgh/index.htm)
- This example also appears in Chapter 3: *Research Training and Career Development*.
- (E) (FIC, NHGRI, NIBIB, NLM)

**Informatics Research Training Programs:** To address the national need for computational scientists competent in biology and medicine, NLM reviewed its University Informatics Research Training Programs and issued a new call for applications. Curricula were updated to reflect current computing needs in clinical translational research and public health. Eighteen 5-year grants, totaling more than \$75 million, for research training in biomedical informatics, were awarded in 2006. Approximately 270 trainees are currently enrolled in these programs.

- For more information, see <http://www.nlm.nih.gov/ep/AwardsTrainInstitute.html>
- This example also appears in Chapter 3: *Research Training and Career Development*.
- (E) (NLM)