Lister Hill National Center for
Biomedical Communications
An Intramural Research Division of the U.S. National Library of Medicine

FY2014 Annual Report

Clement J. McDonald, M.D.
Director

U.S. National Library of Medicine, LHNCBC
8600 Rockville Pike, Building 38A
Bethesda, MD 20894

Lister Hill National Center for Biomedical Communications
FY2014 Annual Report

Table of Contents

Table of Figures

Lister Hill National Center for Biomedical Communications
FY2014 Annual Report

Clement J. McDonald, MD
Director

The Lister Hill National Center for Biomedical Communications (LHNCBC), established by a joint resolution of the U.S. Congress in 1968, is an intramural research and development division of the National Library of Medicine (NLM). Through its biomedical informatics research, LHNCBC develops advanced health information resources and software tools that are widely used in biomedical research and by health IT professionals, health care providers, and consumers. LHNCBC seeks to improve access to high-quality biomedical information for people around the world. It leads programs aimed at creating and improving biomedical communications systems, methods, technologies, and networks and enhancing the sharing and use of information among health professionals, patients, and the general public. The development of next-generation electronic health records (EHRs) to facilitate patient-centric care, clinical research, and public health is an important focus of the LHNCBC as well as an area of emphasis in the NLM Long Range Plan 2006–2016.

The LHNCBC research staff is drawn from many disciplines, including medicine, computer science, library and information science, linguistics, engineering, and education. Teams of people from a variety of backgrounds conduct research that often involves collaborating with other NLM divisions, NIH institutes, and organizations within the Department of Health and Human Services (HHS), as well as with academic and industry partners.

LHNCBC is organized into five major components: the Cognitive Science Branch (CgSB), the Communications Engineering Branch (CEB), the Computer Science Branch (CSB), the Audiovisual Program Development Branch (APDB), and the Office of High Performance Computing and Communications (OHPCC). An external Board of Scientific Counselors meets semiannually to review LHNCBC's research projects and priorities. News and information about LHNCBC research activities are available at http://lhncbc.nlm.nih.gov/.

Biomedical Imaging, Multimedia, and 3D Imaging

The objectives of this research area are to:

- Build advanced imaging tools for biomedical research;
- Create image-based tools for clinical care and medical training;
- Develop multimedia image-text databases that accentuate database organization, indexing, and retrieval; and
- Develop content-based image-retrieval (CBIR) techniques for automated indexing of medical images by image features.

Screening Chest X-rays for Tuberculosis and other Diseases in Rural Africa

In FY2014, we continued our collaborative project with AMPATH (Academic Model Providing Access to Healthcare), an organization supported by the U.S. Agency for International Development (USAID) that runs the largest AIDS treatment program in the world. Through this project, we're conducting imaging research and developing systems to support NIH efforts to improve global health. Our objective is to use our in-house expertise in image processing to clinically screen HIV-positive patients in rural Kenya for lung disease, with a special focus on tuberculosis (TB) and other lung infections prevalent in patients with HIV. We provided AMPATH with lightweight digital X-ray units that are easy to transport in rural areas. The AMPATH staff is taking chest X-rays (CXR) of people and screen them for the presence of disease. This project was selected for HHS Ignite, an initiative of the Department of Health and Human Services (HHS) Innovation Council, and received a cash award that was used to acquire field-deployable equipment. The award identifies projects that have the potential of making radical and positive improvements to the state of the art.

In the past year, one of our X-ray units was mounted onto a truck at the MOI University Hospital in the town of Eldoret in western Kenya. The images from the X-ray units are acquired in a standardized DICOM (Digital Imaging Communications) radiological-image format. Through advances in technologies for Web-based Access to DICOM Objects (WADO) and the implementation of long-range Wi-Fi in western Kenya, images acquired in the field can be stored in the PACS (Picture Archiving and Communications System), a database system used in hospitals to store medical images and housed at the AMPATH building on the hospital grounds in Eldoret.

Figure 1. *(Left) Mobile X-ray unit with generator and on-ramp deployed. (Right) The first patient of the day is prepped for a chest X-ray in Turbo, Kenya in September 2014.*

Because of the lack of sufficient radiological services in western Kenya, we've been focusing our in-house research effort on developing software that automatically screens the CXR images for disease. Our researchers are developing machine-learning algorithms to automatically segment the lungs; detect and remove ribs, heart, aorta, and other structures from the images; and then detect texture features characteristic of abnormalities, which allows us to discriminate abnormal from normal cases. These machine-learning algorithms, which allow computers to learn so they can do a task without being programmed to do it, require large sets of example X-rays. After receiving an internal review board (IRB) exemption, we explored many options for acquiring CXR training sets. We acquired about 400 CXRs from Montgomery County's TB Control Program, 850 from a source in India, 2,000 from a hospital in China, 8,200 from Indiana University, and 250 from an open-source set from Japan. We've made data sets from China and Montgomery County available for access by researchers worldwide.

The usable number of CXRs in our collection is less than the total. This is partly because of the marginal quality of many images and the inclusion of lateral views, which we're not considering yet. To create a training set, we manually validated and annotated the images, and in FY2013, we completed this process for the images from Indiana University. To ensure the privacy of the X-rayed subject, we reviewed every image to verify that it showed no identifying information. Also, we eliminated images that included noisy data, such as the stray wires or tubes that are common in hospital-based radiological imaging, since they tend to confuse the classifiers.

Using these X-rays for training and testing, we developed algorithms for detecting relevant anatomy in the chest regions such as lungs, ribs, and the heart. To ensure the acquisition of high-quality images, we are using rib structures to develop other algorithms that detect the planar and positional rotation of the patient because such rotation can lead to incorrect diagnosis. Being able to detect the heart allows us to separate it from the lung regions, then we can also detect abnormalities such as cardiomegaly (enlarged heart) that can be a precursor to congestive heart failure.

During FY2014, we were consulted by the National Institute of Allergy and Infectious Diseases (NIAID) about adapting our algorithm to detect pediatric tuberculosis. As a first step, we adapted the lung-segmentation algorithm to include pediatric lung shapes and found that it is reasonably accurate. We need a much larger training and test set to provide meaningful measures of performance quality. In 2014, we speeded up our lung-segmentation method, which can now segment lungs in tens of seconds rather than minutes. Radiologists from the NIH Clinical Center, Yale University, and the University of Missouri in Columbia helped us by annotating pathology in some of our images. We used these annotated images to train our Support Vector Machine (SVM)–based classifier, which uses several features extracted from the X-rays as input, such as histograms of intensity, gradient magnitude and orientation, shape, and curvature. On the basis of these input features, the SVM returns a confidence value, allowing an operator to inspect cases about which the classifier is uncertain. We also compared the performance of the algorithm classifier with that of human experts. We found that they perform similarly (87 percent accurate), but the classifier tends to be more sensitive, yielding nearly twice as many false positives. While not ideal, this oversensitivity does prevent overlooking X-rays that show disease, and it's useful in a resource-constrained setting. We are continuing to research methods for advancing classifier performance by sampling image patches.

3D Informatics for High-Resolution Microscopy

We continue to address problems encountered in the world of three-dimensional and higher-dimensional, time-varying imaging through our 3D Informatics (3DI) and Molecular Visualization programs.

Throughout FY2014, we continued our collaboration with NCI's Laboratory for Cell Biology and with teams within LHNCBC to visualize and analyze complex 3D volume data generated through dual-beam (ion-abrasion electron microscopy) and cryo-electron tomography. This work applies high-performance computing and audiovisual didactic development to data from life sciences research related to the detection and prevention of cancer and infectious diseases. The resulting visuals have provided insights about the character of several immunological cells, cell structures, and the cells' interaction with pathological viruses, including HIV. In collaboration with NCI, we developed a 3D illustration of HIV-1 transmission, which was featured on the cover of the *Journal of Virology*.

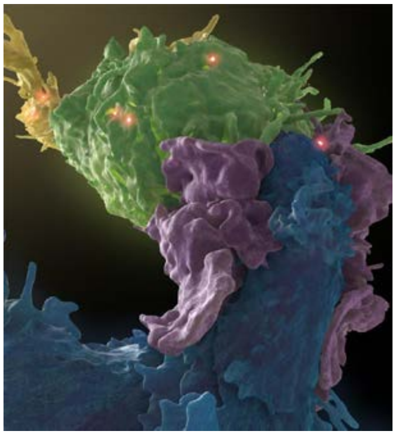


Figure 2. *Uninfected blue and purple T-cells push through their environment to reach the green and yellow HIV infected T-cells in an image featured on the cover of the Journal of Virology. The work describes morphological differences in the virological synapses of a T-cell cell line (Jurkat cells) and primary human CD4 T-cells from peripheral blood. (*J Virol. *2014 Sep;88(18):10327-39.)*

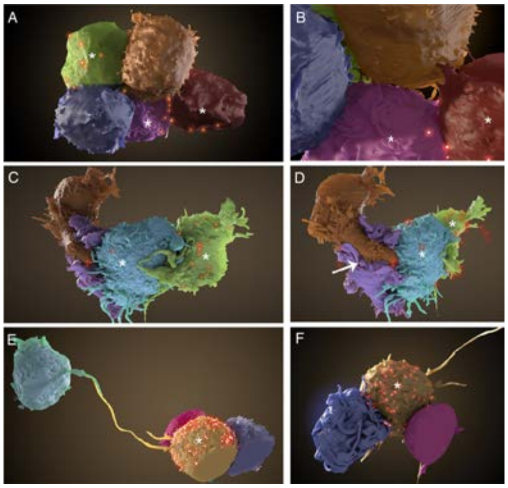


Figure 3. *Segmentation of volume cell data shows morphological (shape) differences between common lab grown T-cells (top row) and T-cells from infected host. These differences*

*could have an impact on vaccine research. Red spheres represent surface bound HIV. (*J Virol. *2014 Sep;88(18):10327-39.)*

NLM worked with NCI to create a 3D animation of the HIV infectious spike. The 3D animation revealed a formerly undescribed twisting motion that affects binding to target membranes. Another collaboration with NCI clarified the structural mechanisms of glutamate receptor activation and desensitization in a *Nature* paper. We also used 3D-imaging software to translate data from cutting-edge NCI microscopy into illuminating visuals that led to new discoveries about influenza, including a publication about the feasibility of a universal influenza vaccine.

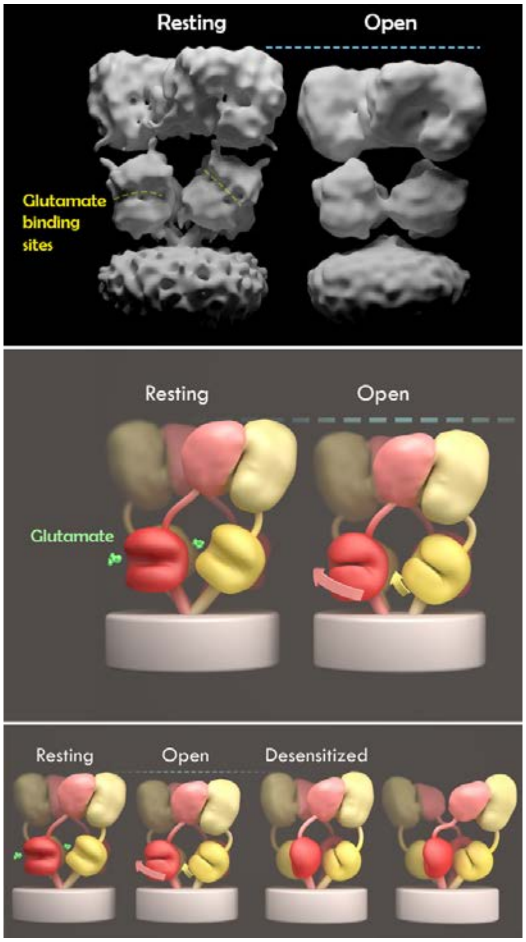


Figure 4. *Stylized 3D model, we developed in collaboration with NCI show the discovery of relatively large molecular movements in the desensitized state of the glutamate receptor that are much harder to interpret directly from the data.*

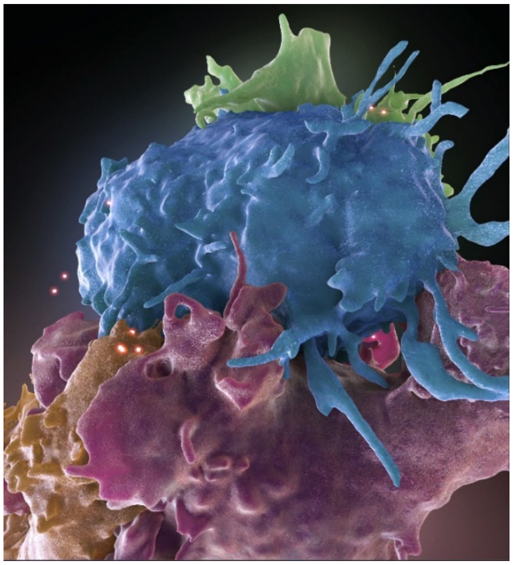


Figure 5. *LHNCBC created this image, which features HIV-infected T-cells. (*J Virol. *2014 Sep;88(18):10327-39.) The* New York Times *requested and published it to illustrate a virus in their article entitled "Ebola and the Vast Viral Universe," published October 28th, 2014.*

We continued our commitment to processing data collected through transmission electron tomography. We successfully published the results of software-development research that uses graphics processing units (GPUs) for high-performance computing for sub-volume averaging and reconstruction. (The image data collected are equivalent to the serial sections of a CT scan. These sections can be put back together to "reconstruct" the original "volume." Any mathematics performed on part of the "volume" is called "sub-volume.") We're working now to develop emerging methods into mature software that can extend this work on sub-volume reconstruction to single-particle microscopy. If successful, the software will expand our impact to a wider range of molecular targets in systems biology.

This past year, we also helped supervise segmentation efforts for data from ion-abrasion electron microscopy in a study of the disposition of malaria pathogens (that is, where they wind up) in normal human blood cells. These 3D data are collected with resolutions of 3 nanometers across a field of view that can encompass entire cells. We're evaluating the impact of heterogeneous sickle cell hemoglobin on the infiltration, geometry, and spatial relationships of the pathogens in affected erythrocytes. Other work in related areas includes the segmentation and study of vaccinia viruses in normal and mutant human liver cells.

Content-Based Image Retrieval (CBIR)

CBIR is an active research area in the imaging research community since many of its tools and techniques find application in systems for image indexing, search, and retrieval. A goal of this research is to find images in repositories or the published literature that are visually or semantically similar to an image or text query. For example, one chest X-ray might be visually similar to another, but semantic similarity lies in finding another chest X-ray with the same lung disease.

We have developed several practical systems and tools that rely on CBIR research. For instance, our Open-i system allows users to access 1.77 million figures from medical journals, including photographs, clinical images, charts, and other illustrations. People can sort search results based on different types of images, starting with "regular" and graphical images. Graphical images are further categorized as diagrams, statistical figures, flow charts, and tables. Regular images are further categorized as X-ray, ultrasound, computed tomography (CT), magnetic resonance imaging (MRI), and other modalities. We use more than 15 image modalities to classify the images with our Support Vector Machine (SVM)–based framework. These modalities are image features, such as color, texture, and shape. We successfully used our modality classification system during international ImageCLEFmed competitions, and we incorporated it into Open-i. The Open-i project is described in more detail below.

Authors place arrows on figures to highlight important regions. To help people find the images they need more quickly, we extract specific regions of interest (ROIs) within images. By analyzing the image layout, we've improved the performance of algorithms we developed for detecting and extracting the arrows.

Authors place arrows on figures to highlight important regions. We're developing additional algorithms to retrieve images that are more relevant to the query, whether the query is submitted as text, a photo, or a combination. To retrieve relevant images, we first determine how the visual data correspond to the concepts in the query text. Using a method that divides the image into tiles, or image patches, we group each image patch with other similar-appearing patches across all images in the database. We develop a correspondence between representative image patches selected, frequently occurring groups, and key biomedical concepts in the accompanying text. We then apply machine-learning algorithms to extend this labeling to all patches in various groups so that every image patch has a text label associated with it. Because image patches are derived from images, every image is now transformed from a pixel-based representation into an image document where patches are replaced with these group labels.

The advantage of this approach is that not only are we able to map text queries to visual data, we're also able to apply fast, traditional, text-based information-retrieval techniques to image retrieval. We're working toward extending this idea to text-phrase retrieval strategies to find images with relevant local regions of interest. Recent advances in this area have led to a retrieval relevance of 75 percent for image retrieval using text queries such as, "Find lung CT images with ground-glass opacity."

We use spatial layout of pixel intensities within the image to eliminate regions that are not likely to be arrows, then we apply structural information about the arrow shape to identify candidate arrow regions. We used to use Markov Random Field (MRF) models to recognize arrow-type pointers with a precision of 85 percent and recall of 82 percent. Our new algorithmic approach has improved our precision to 94 percent and recall to 87 percent. Furthermore, our algorithm automatically detects whether the arrow is of a lighter or darker color compared with the background, which enables us to successfully apply it to a wide variety of images.

We're also developing a correlation between image ROIs and key biomedical concepts that appear in neighboring text, such as figure captions or other text describing the image content. Image features used to index the entire image may aggregate the details in specific ROIs.

Another example of the role of CBIR in our work is in our development of CervigramFinder, a research tool that automatically indexes and enables the retrieval of uterine cervix images (cervigrams) by shape, color, and texture features. Being able to search efficiently by image features is a significant step toward locating records in large databases of cervigrams and patient data, such as NCI's Guanacaste and ASCUS-LSIL Triage Study (ALTS) databases, containing a total of 100,000 cervigrams. We've made advances in this area by developing algorithms that use values from several fields from the patient record, such as the woman's age and HPV-infection along with image features from the colposcopic exams. We're using these data to develop a model for predicting the likelihood that a patient will progress to more severe forms of HPV-based uterine cervical infections, including precancerous cells. In FY2014, we worked with Lehigh University to further develop this capability in the Multimodal Cervix Classification System (MCCS), which has the goal of accepting as input a uterine cervix image, with additional optional patient information such as HPV status, and producing as output a classification of the image into a low-risk (Normal/CIN1) or high-risk (CIN2/CIN3) category.

CBIR is also allowing us to improve the use of chest X-rays in an automated approach to detecting tuberculosis and other pulmonary diseases, which could be very useful in resource-poor countries. We've developed algorithms to automatically detect ribs, aorta, and other structures and to segment lung areas. Research continues into extracting texture features to classify lungs as normal or abnormal using SVM classifiers. This project is described in more detail above.

In FY2014, in collaboration with NIAID, we developed an early prototype for automatically counting infected cells in blood films to assist malaria diagnostics, including cell detection and classification, and we finished implementing an image-processing method for measuring closing rates of cell-migration assays to support basic research of critical pathways for tissue repair. This method relies on many other methods used in the CBIR research area, including identifying cell boundaries and determining which cells are infected. Our malaria-imaging research received the Best Abstract Award at the Third Annual Seminar on Molecular Imaging of Infectious Diseases.

Other areas of our research include using distributed computing and GPUs for computer-intensive CBIR tasks, especially image segmentation. Through our collaboration with Texas Tech University, for example, we developed a method that uses GPU processing power for interactively following challenging object boundaries, such as the separation between the epithelial and nonepithelial tissue in histology slides of the uterine cervix. To support early detection and improve healthcare outcomes for people with cervical cancer, we plan to use these segmented epithelial regions to train doctors to detect various stages of precancer.

Imaging Tools for Biomedical Research and Education

In collaboration with cancer researchers and clinicians who conduct screening and diagnostic tests for cervical cancer, LHNCBC conducts research and develops innovative biomedical imaging tools.

The American Society for Colposcopy and Cervical Pathology (ASCCP) continued to use one of our image-based systems, the Teaching Tool, to assess the knowledge and skills of colposcopy professionals. More than 130 resident programs in Obstetrics/Gynecology and Family Practice at more than 100 universities and other premier institutions, such as the Mayo Clinic, Baylor College of Medicine, Duke University, and the Tripler Army Medical Center, have used the tool. Since we first released the Teaching Tool in May 2010, these programs have administered more than 2,000 individual online exams (the Residents' Assessment of Competency in Colposcopy Exam (RACCE)); in addition, the ASCCP has used the Teaching Tool to administer the more-advanced Colposcopy Mentorship Program (CMP) exam to more than 370 practicing colposcopists.

In 2014, we worked with the ASCCP to build more content for these exams by supporting the collection of expert colposcopist interpretations and biopsy regions for a set of 415 uterine cervix images. We plan to add the subset of these images on which the experts reached consensus opinion on quality, visual diagnosis, and biopsy region to the pool of image-based questions used by the RACCE and CMP exams. We also worked with the ASCCP to take steps toward the creation of an additional exam, the Colposcopy Recognition Award (CRA) exam, which is to test the highest level of expertise recognized by the ASCCP. We put a first, trial version of the CRA online in 2014 for ASCCP evaluation and anticipate bringing the final CRA online in 2015, depending on when the ASCCP reaches final consensus on content.

Our Boundary Marking Tool (BMT), another imaging program, was used extensively in FY2014 for the data collection carried out for the work with the ASCCP. The 415 uterine cervix images mentioned above were distributed to more than 20 colposcopy experts in a randomized manner, with each image evaluated by three experts. Three rounds of data collection were carried out, and in August, the BMT was used at a working meeting with LHNCBC staff and ASCCP experts to view the collected expert evaluations (that is, judgments on image quality, visual diagnosis, and biopsy regions) and reach expert consensus. The group ultimately selected 123 images for use in future Teaching Tool exams.

We also continued research into methods for the computerized analysis and classification of cervical tissue using images collected and annotated by pathologists at the University of Oklahoma Health Sciences Center. This work, with collaborators at the Missouri University of Science and Technology, includes applying our algorithm to carry out nuclei segmentation within the epithelial regions of the tissue, deriving image features based on this segmentation, and using the features in Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) classifiers for automated classification of the epithelium into classes of normal or various grades of abnormal (CIN1, CIN2, and CIN3). In the latest experimental work, we achieved classification accuracy of 88 percent in the automatic labeling of the epithelium as normal or one of the three classes of abnormal, using the LDA classifier with 27 image features to label each of 61 epithelium images that a pathologist with expertise in the histology of the uterine cervix had labeled.

In FY2014, we installed and began the trial use of technology developed in collaboration with Texas Tech University to view, manually segment, and then do tissue classification of regions in uterine cervix histology images through a Web-based system. This system includes the Intelligent Scissors algorithm (well known in the image-processing community) to assist and speed up the segmentation process and will take advantage of graphics processing unit (GPU) hardware for some of the image-processing computations, if a GPU is available on the user's local system.

We began implementation of a new architecture for our Multimedia Database Tool to move that system from a Java client application to a system with a lightweight client running in a Web browser, to make the system access as easy as possible to the widest group of users. We also began re-implementing the server side of the system with the latest Java technology using Java Server Faces in particular, which we expect will speed up our development and make maintenance and future modification more tractable.

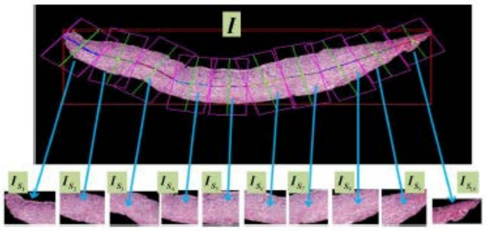
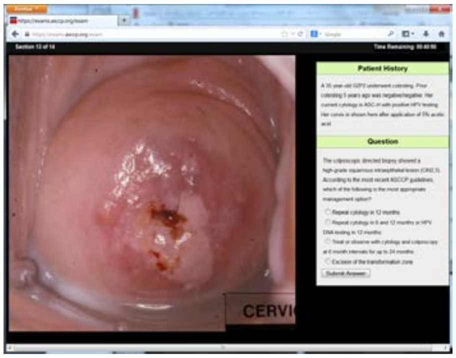


Figure 7. LHNCBC automatically processed this image to find the center line ("medial axis") for a segment of uterine cervix tissue, and then to divide the tissue into multiple segments for detection of cervix disease. (A collaboration among the Communications Engineering Branch, Missouri University of Science & Technology, and the University of Oklahoma Health Sciences Center.)



Patient History

A 35 year-old G2P2 underwent cotesting. Prior cotesting 5 years ago was negative/negative. Her current cytology is ASC-H with positive HPV testing. Her cervix is shown here after application of 5% acetic acid.

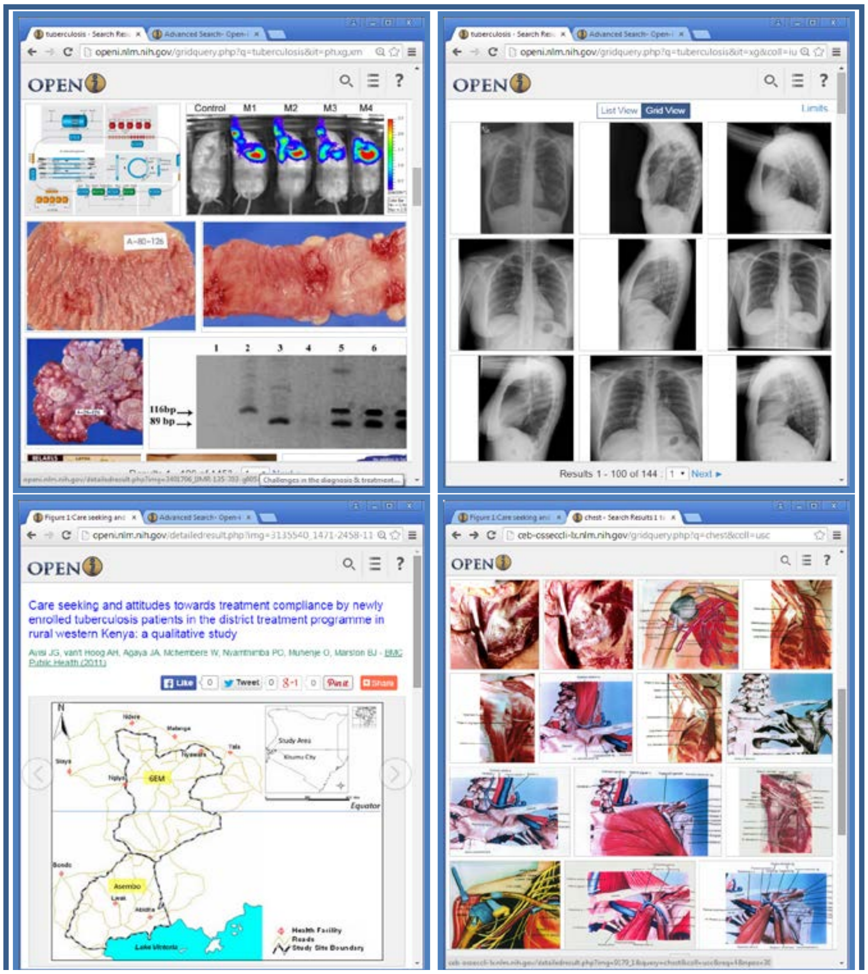
Question

The colposcopic directed biopsy showed a high-grade squamous intraepithelial lesion (CIN2,3). According to the most recent ASCCP guidelines, which of the following is the most appropriate management option?

○ Repeat cytology in 12 months
○ Repeat cytology in 6 and 12 months or HPV DNA testing in 12 months
○ Treat or observe with cytology and colposcopy at 6 month intervals for up to 24 months
○ Excision of the transformation zone

Figure 6. The American Society for Colposcopy and Cervical Pathology (ASCCP) continues to use LHNCBC's Teaching Tool software to administer online, nationwide proficiency exams in the field of colposcopy. This figure shows a representative question in how to manage a patient, given clinical history and a current colposcopy image of the uterine cervix.

A picture is worth a thousand words, especially in medical research and clinical practice! Most people can understand complex biomedical concepts more easily if they are presented visually: through radiographic images, photographs of organs, sketches, graphs, or charts. This idea motivates our project, part of the LHNCBC Clinical Information Systems effort, which exploits ongoing research in both natural language processing and content-based image retrieval by processing and indexing images based on both text and image features. We developed the Open-i system for finding images and figures in published literature or other sources. It gives medical professionals and the public access to images contained in biomedical articles that are highly relevant to their query, as well as a summary of the articles' take-away messages. Users may search by text queries as well as by example images. They can filter images by type (e.g., X-ray, graph), filter journals by clinical specialty, and rank papers by clinical task (e.g., treatment). This ability assists clinical decision support, biomedical research, and education.



Care seeking and attitudes towards treatment compliance by newly enrolled tuberculosis patients in the district treatment programme in rural western Kenya: a qualitative study

Ayisi JG, van't Hoog AH, Agaya JA, Mchembere W, Nyamthimba PO, Muhenje O, Marston BJ - BMC Public Health (2011)

OPEN
openi.nlm.nih.gov/gridquery.php?q=tuberculosis&it=ph,xg,xm
openi.nlm.nih.gov/gridquery.php?q=tuberculosis&it=xg&coll=iu
openi.nlm.nih.gov/detailedresult.php?img=3135540_1471-2458-11
ceb-ossecdl-lx.nlm.nih.gov/gridquery.php?q=chest&coll=usc
Control M1 M2 M3 M4
1 2 3 4 5 6
116bp
89 bp
List View
Grid View
Limits
Results 1 - 100 of 144
Next ▸
Like
Tweet
Pin it
Share
N
Study Area
Kisumu City
GEM
Asembo
Lake Victoria
Equator
Health Facility
Roads
Study Site Boundary

__Figure 8.__ Two collections added to NLM's biomedical image search engine Open-I allow scientists and the public to view images in PubMed Central articles related to their searches (upper left), drill down to details about images of interest (bottom left), and view related images in the collection of chest radiology reports provided by Indiana University (upper right) and anatomy illustrations provided by University of Southern California (bottom right).

Open-i was released to the public in 2012 and is the first production-quality system of its kind in the biomedical domain. The system enables users to search and retrieve medical citations from 517,715 open-access articles in PubMed Central® (PMC). In FY2014, we increased the collection of images in Open-i to 2.2 million images from PubMed Central articles (from 1.3 million in FY2013) and to 7,470 from Indiana University radiology reports.

The quality of the information delivered by Open-i has been evaluated in international medical-image-retrieval "Cross Language Evaluation Forum" competitions (ImageCLEF), in which the system consistently ranked among the best. For example, the system demonstrated the best retrieval results in a 2013 ImageCLEF that attracted participants from academia, industry, and clinical settings.

The demand for Open-i services grew 26 percent in 2014. We upgraded the system architecture and the user interface to be consistent across desktop computers and mobile devices, and increased the number of images shown per page from 20 to 100.

The system is available 24/7, and it can handle more than 20,000 interactions per day in real time. We have also added an application programming interface (API) that provides batch-retrieval services for researchers who need access to images. For example, in 2014, Pittsburgh researchers used the Open-i API to build a visual ontology for carcinoma biomarkers and add illustrations to their oral squamous cell carcinoma dataset. Colorado researchers downloaded the Indiana X-ray collection from Open-i and plan to map image features to concepts. We capture the most popular searches and display the top images for the three most popular searches on the Open-i home page. This new feature allows users to quickly navigate to the results of popular searches that were trending in 2014, such as for Ebola, keloids, pulmonary tuberculosis, and lung tumor.

According to data from our internal logs, Open-i serves up to 20,000 distinct users a day. Of these, per Google Analytics' estimates, about 6,500 are unique visitors, and the rest are bots from search engines.

In other work this past year, we explored deep-learning methods for building a visual structural framework for organizing information called a visual ontology. We began to explore the use of these methods to automatically detect the content in the medical image. We also developed classifiers to identify photographic images containing some form of "tissue" in the image. In this context, "tissue" is defined as images containing dermoscopic, endoscopic, ophthalmic, dental and oral, and similar images. We developed methods that detect such images with high accuracy from other general photos in the Open-i collection and from other clinically less important data. The evolution of Open-i continues to be supported by research in these key areas:

- Information retrieval and search engine optimization;
- Text summarization;
- Representing images with text strings;
- Combining global and local representations of image features;
- Improving methods for automatically segmenting multi-paneled illustrations into single images and partitioning their captions to correspond to those single images; and
- Improving methods for extracting pointers (arrows, arrowheads, symbols) within images to identify regions of interest that could be more relevant to a query than the entire image would be.

Computational Photography Project for Pill Identification (C3PI)

Launched in September 2010, the Computational Photography Project for Pill Identification (C3PI) is developing an authoritative, comprehensive, public digital-image inventory of the nation's commercial prescription, solid-dose medications. Working with expert consultants, we are creating a collection of photographs of prescription tablets and capsules, confirming that the images match the description of the medication, and developing and matching the images of the samples to relevant metadata (such as size descriptions, dimensions, color, and the provenance of the sample).

In FY20144, we increased our online collection to more than 808,000 images of more than 3,000 pill (solid-dose pharmaceutical) samples from more than 150 manufacturers and distributors. The team generates high-resolution, high-quality pill images from a variety of lighting conditions. The long-term goal is to develop computer-based tools to help identify an unknown medication based on image-matching algorithms.

These images are freely available online directly via an API or through an interactive Web search via NLM's Pillbox and RxNav sites. Through two online repositories — http://rximage.nlm.nih.gov (which links to RxNav) and http://splimage.nlm.nih.gov — we can distribute images of oral, solid-dose medications to the public and to pharmaceutical manufacturers, respectively.

In FY2014, we began a collaborative project with the Food and Drug Administration (FDA) to extract drug-drug interactions from drug labels to support FDA's structured product labeling (SPL) indexing initiative.

The Visible Human Project

The Visible Human Project image datasets serve as a common reference for the study of human anatomy, a set of common public domain data for testing medical imaging algorithms, and a source for the anatomical data needed to model human structure and physiology. These datasets are available through a free license agreement with the NLM.

We distribute them in their original or in PNG (Portable Network Graphics, a raster graphics file format that supports lossless data compression so the images retain their original resolution) format to licensees over the Internet at no cost and on DVDs for a duplication fee. Almost 44,000 licensees in 64 countries are applying the datasets to a wide range of educational, diagnostic, treatment-planning, virtual-reality, and virtual-surgery uses, in addition to artistic, mathematical, legal, and industrial uses. In FY2014, we continued to maintain two databases for information about how people are using the Visible Human Project: one for information about the license holders and their intended use of the images, and the other for information about the products the licensees provide NLM as part of the Visible Human Dataset License Agreement. More than 1,000 newspaper, magazine, and radio pieces have featured the Visible Human Project since we released the first dataset in 1994.

Insight Tool Kit

The Insight Toolkit (ITK) is a public, open-source algorithm library for segmenting and registering high-dimensional biomedical-image data. In FY2014, we released the official software version ITKv4.6.0, which includes a collection of more than 1,500 filters and algorithms for medical image processing. More than 845,000 lines of openly available source code comprise ITK, and with it, people can access a variety of image-processing algorithms for computing segmentation and for registering high-dimensional medical data. ITK runs on Windows, Macintosh, and Linux platforms, so it can reach across a broad scientific community. It is used by more than 1,500 active subscribers from 40 countries. A consortium of university and commercial groups, including our intramural research staff, provides support, development, and maintenance of the software.

In FY2014, ITK's project manager, Terry Yoo, PhD, received the Hubert H. Humphrey Award for Service to America, "for leadership in open science through ITK as the public software standard for 3-dimensional biomedical-image analysis and for all the patients who have benefited."

We also updated SimpleITK (sITK) to broaden accessibility of ITK to the Python programming language community. Developed primarily by programmers at NLM, sITK is a simplified layer built on top of ITK. It is intended to facilitate ITK's use in rapid software development and education through the support of scripting languages, primarily Python. The sITK interface conceals the structural and design complexities of ITK, enabling more straightforward, procedure-based programming. Designed to be an interpreted scripting system, sITK supports a typeless, polymorphic data model, thus simplifying the use and expression of ITK in image-analysis education. ITK remains an essential part of the software infrastructure of many projects across and beyond NIH. Since its inception in 2004, the Harvard-led National Alliance of Medical Image Computing (NA-MIC), an NIH Roadmap National Center for Biomedical Computing (NCBC), has adopted ITK and its software-engineering practices as part of its engineering infrastructure. Though the 10-year funding for the NIH Roadmap is now complete, ITK remains integrated with 3D Slicer, the open-source software surgical-planning workstation developed by NA-MIC. ITK also serves as the software foundation for the Image Guided Surgery Toolkit (IGSTK), a research and development program sponsored by the NIH National Institute for Biomedical Imaging and Bioengineering (NIBIB) and executed by Georgetown University's Imaging Science and Information Systems (ISIS) Center. IGSTK is pioneering an open application programming interface (API) for integrating robotics, image guidance, image analysis, and surgical intervention. In FY2014, the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) awarded to ITK a Small Business Technology Transfer (STTR) grant to build a virtual-surgery system to treat children with craniosynostos. The new project will build on the experience of the development of IGSTK, ITK, and 3D Slicer.

International software packages that incorporate ITK include OsiriX, an open-source, diagnostic radiological-image viewing system available from a research partnership between UCLA and the University of Geneva, and the Orfeo Toolbox (OTB) from the Centre Nationale D'Etudes Spatiales, the French National Space Administration. Beyond the support of centers and software projects, the ITK effort has influenced end-user applications through supplementing research platforms such as the Mayo Clinic's Analyze and University of Utah's Scientific Computing and Imaging Institute's SCIRun, and through developing a new release of VolView, free software for medical volume image viewing and analysis.

Over the past year, we have seen innovation in and adoption of both ITK and sITK. The NLM software initiative continues to have impacts in research and education. At least three educational initiatives have used sITK to teach medical volume image viewing and analysis, including courses at Rensselaer Polytechnic Institute (RPI), the University of Iowa, and the Imperial College of London. In keeping with the principles of open science, the tutorial and educational materials are all made available without cost or license restrictions online.

3D Printing

NLM has been contributing to the NIH 3D Print Exchange, an online resource for searching, browsing, and downloading 3D printable files related to medicine and biosciences. This consortium of three NIH ICs — NIAID, NLM, and NICHD — received funding and backing from the HHS Ignite and HHS Ventures initiatives of the HHS IDEA Lab. The 3D Print Exchange officially launched at the White House Maker Faire in FY2014.

Designed to supply biomedical shape files for education and research to a growing audience of users worldwide, the exchange provides public software built from NLM's Insight Toolkit (ITK) and Kitware's Visualization Toolkit (VTK) to automatically generate printable models from X-ray CT data. The project's data repository is officially online, and a national advisory board is assisting with its oversight, testing, and guidance.

The 3DI group also continued to investigate the use of rapid-prototyping technologies in radiology with partners at NIAID. We analyzed the X-ray attenuation characteristics of the 3D-printing materials available at NIH and are presently evaluating the use of contrast agents as printing materials to vary the appearance of the 3D models. A new set of models is under development, including dosimetry models from CT scans of small animals.

Natural Language Processing and Text Mining

Medical Article Records System (MARS)

NLM's flagship database, MEDLINE®, contains more than 21 million bibliographic records for articles from more than 5,600 biomedical journals. To meet the challenge of producing these citations in an affordable way, researchers at LHNCBC develop automated techniques to extract bibliographic data, such as abstract, author names, and affiliations, from both scanned-paper and online journals.

While the bulk of citations now comes to NLM directly from publishers (in XML format), nearly 707 journals provide citations in paper form only. These papers are processed by the MARS, which we launched in 1996. MARS combines document-scanning, optical character recognition (OCR), and rule-based and machine-learning algorithms to extract citation data from paper copies of medical journals for MEDLINE. The stages in this automated process are:

- Segmenting page images into text zones,
- Assigning content labels (title, author names, abstract, grant numbers, etc.) to the zones, and
- Pattern matching to identify the specific entities in each zone.

We manage and continually improve the MARS system. For example, we have introduced two new features that correct and verify the final output: (1) the ability for the Edit operators to correct errors made by the automated zoning and labeling process, and (2) a Web-based user-interface design for Edit and Reconcile operators. We have completed the software implementation and the integration test for both features, which are now being deployed to the production system used by NLM Library Operations.

We have also developed a system (Publisher Data Review System, or PDRS) to supplement the citation data that publishers send in electronically (in XML form), since these often contain errors or missing fields. Examples of such missing items required in MEDLINE citations are databank accession numbers (e.g., for items in GenBank and similar databases), NIH grant numbers, grant support categories, investigator names, and information about the links between articles and the comments submitted in response to them. The capture of investigator names can be especially difficult because some articles contain hundreds of such names, and capturing the articles that "comment on" another paper (usually an editorial or a review article) requires operators to open and read other articles related to the one being processed. PDRS subsystems are based largely on machine-learning algorithms such as the support vector machine, or SVM. The PDRS went into production in early FY2012 for open-access articles in NLM's PubMed Central and Figure 9 shows the total number of online articles being processed monthly by PDRS in 2014.

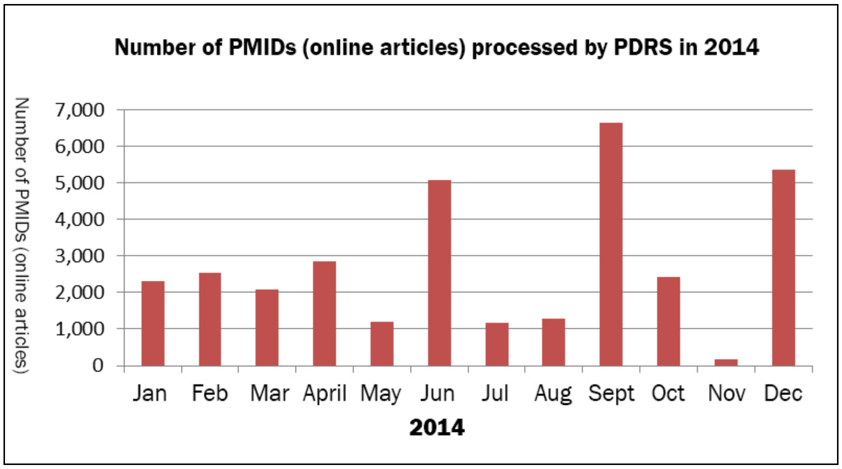


Figure 9. *Number of PMIDs (online articles) processed by PDRS in 2014.*

To extend automated indexing data extraction to *all* online journals on publishers' sites, including the journals with restrictive copyrights, we're developing IMPPOA (In-Memory Processing for Publisher Online Articles). This is a system based on the PDRS platform and its machine-learning algorithms, but it retains articles (for processing) for only a very short time in random access memory (RAM). The IMPPOA system:

- Provides data missing from the XML citations sent in directly by publishers,
- Corrects errors in publishers' data by extracting data from the articles on their sites and comparing these with the data sent to NLM, and
- Extracts data from articles for which publishers do not send in citations at all.

Because this new system avoids downloading the articles to a disk drive, we expect IMPPOA to eliminate publishers' concerns about copying articles into an external system disk.

The systems outlined above rely on underlying research in image processing and lexical analysis that also enables the creation of new initiatives for applying these techniques, such as the ACORN project (also known as the Automatically Creating OLDMEDLINE Records for NLM project), described below.

Automatically Creating OLDMEDLINE Records for NLM (ACORN)

NLM has a long-standing goal to expand MEDLINE to include all bibliographic records going back to 1879, when *Index Medicus* was first introduced. The earliest citations exist only in printed paper indexes, and Library Operations (LO) has collected many of these with considerable manual effort. To automatically extract citations from these paper indexes, we designed the ACORN system, which combines scanning, image enhancement, optical character recognition (OCR), image analysis, pattern matching, and related techniques. The challenges are significant, including the fact that these indexes used old typefaces and fonts and a mix of different languages—all contributing to highly inaccurate OCR results. To overcome these problems in one of the indexes (Quarterly Cumulative Index Medicus, or QCIM), we developed a novel pattern-matching technique that automatically finds and compares two versions of every citation, one from the subject listing and the other from the author listing. This minimizes the OCR errors encountered in each version. In addition, our system searches MEDLINE to find records that may already exist, to avoid duplications. Figure 6 shows the context-level diagram of the ACORN system.

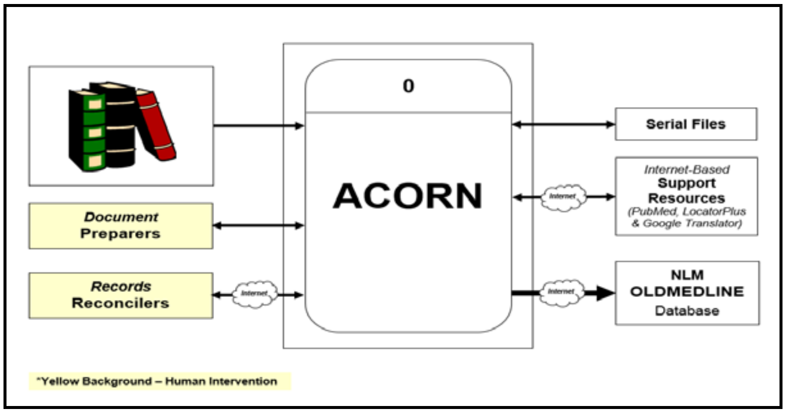


Figure 10. *The ACORN Context-Level Diagram.*

ACORN consists of three main graphical user interface (GUI) subsystems: (1) Scanning and Quality Control, (2) Document Profile Creation and Journal Title Extraction, and (3) Web-based Reconciliation (that is, when operators verify the extracted information). We completed and delivered the first subsystem to NLM LO in April 2013, and by the end of 2014, the LO staff had scanned all 60 QCIM volumes and had completed quality-control checks of QCIM volumes from 60 to 39. We are still developing the other two subsystems, and plan to release the second one in June 2015.

Indexing Initiative

The Indexing Initiative (II) project investigates language-based and machine-learning methods for the automatic selection of subject headings for use in both semiautomated and fully automated indexing environments at NLM. Its major goal is to facilitate the retrieval of biomedical information from textual databases such as MEDLINE. Team members have developed an indexing system, Medical Text Indexer (MTI), based on two fundamental indexing methodologies. The first of these calls on the MetaMap program to map citation text to concepts in the Unified Medical Language System® (UMLS®) Metathesaurus, which are then restricted to Medical Subject Headings (MeSH®). The second approach uses the MeSH from the PubMed-related articles. Results from the two basic methods are combined into a ranked list of recommended indexing terms, incorporating aspects of MEDLINE's indexing policy in the process.

NLM Library Operations (LO) MEDLINE® citations indexers regularly (and increasingly) use the MTI system. To facilitate their indexing, MTI provided recommendations for 708,911 articles in FY2014 as an additional resource available through the Data Creation and Maintenance System (DCMS). Indexers consulted the MTI recommendations for 62.98 percent of the articles in FY2014, compared with 59.34 percent in FY2013 and 57.97 percent in FY2012. MTI also provides indexers with an option to select MeSH heading-subheading pairings for some of the MTI recommendations based on our subheading attachment software. The MTI developers have also created special filtering for MTI to assist in the indexing of the NLM History of Medicine book collection and for general cataloging. Due to its success with certain journals, MTI has been designated as the first-line indexer (MTIFL) for those journals. As a "first-line" indexer, MTI indexing is subject to human manual completion and review. The number of MTIFL journals will grow gradually and should prove to be a time and money saver for NLM.

In FY2014, we doubled the number of new journals for the third year in a row by adding 110 to the MTIFL program (for a total of 230, compared with 120 in FY2013 and 45 in FY2012), which included 29,956 articles (compared with 9,771 in FY2013 and 4,205 in FY2012). We are collaborating with LO to evaluate how well MTI is performing on indexing the journals that are already part of the MTIFL program by computing standard information-retrieval measures (recall, precision, and f-measure) and then comparing MTI's indexing recommendations with the final, official MEDLINE indexing. We also work with LO to identify future MTIFL journal candidates.

Our collaboration with LO and experience gained by participating in the BioASQ challenges – for biomedical semantic indexing and question answering (http://www.bioasq.org/) – have helped increase performance of the MTI system to 58.07 percent (from 56.33 percent in FY2013 and 54.81 percent in FY2012).

In FY2014, MTI provided the primary baseline for the second year of the international BioASQ challenge. The aim of the challenge is to make biomedical text more accessible to researchers and clinicians. The MTI indexing results provided one of the baselines used in the “large-scale online biomedical semantic indexing” part of the challenge, which is designed to parallel the human indexing currently being done at NLM. The II team provided help and guidance in developing the list of journals used in the challenge, as well as the baseline results. MTI will also provide a baseline for the third year of the BioASQ challenge in 2015.MetaMap is a system for identifying concepts within text documents. It is a critical component of the MTI system and is also used worldwide in bioinformatics research. In FY2014, users downloaded about 2,100 copies of MetaMap and 1,200 copies of the Java API and Unstructured Information Management Architecture (UIMA) wrapper.

Digital Preservation Research (DPR)

The long-term preservation of documents in electronic form, both born-digital and scanned from paper, is a mandate for NLM, as it is for other major libraries and archives. The goal of the LHNCBC DPR project is to investigate and implement techniques for these key preservation functions: automatically extracting metadata to enable future access to the documents, ingesting the documents and metadata into a storage system, and conducting knowledge discovery on the archived material. To provide a platform for this research, we built and deployed a System for Preservation of Electronic Resources (SPER). SPER builds on open-source systems and standards (e.g., DSpace, Resource Description Framework (RDF)) while incorporating in-house-developed modules that implement the preservation functions listed above.

In 2014, we completed the preservation of a historic medico-legal collection of early 20th-century court documents acquired from the FDA. These are “notices of judgment” issued by courts against companies that were indicted for misbranding or adulterating foods, drugs, or cosmetics. They offer insights into legal and governmental history dating from the 1906 Food and Drug Act and illustrate regulatory impacts on public health. NLM curators have been using SPER to preserve the FDA documents, numbering 67,000 in total. In 2014, we completed the processing of the fourth and final set of court case summaries. This set comprised 30,700 notices of judgment, published between 1940 and 1966 and related solely to adulterated or misbranded foods. The documents were added to the archive, and, along with metadata, they are publicly accessible at an NLM Web site and are being used by researchers. The second collection, from NIAID, is a set of conference proceedings of the U.S.-Japan Cooperative Medical Science Program (CMSP) Cholera and Other Bacterial Enteric Infections Panel, an international program conducted over a 50-year period from 1960 to 2011. For this collection, our activities include: (1) building a full repository for 2,800 research articles on cholera and 8,000 references on CMSP participants such as authors, panelists, attendees, and study section reviewers, followed by (2) developing a portal where the public can go to search for research articles, institutes, and authors. To support these activities, we developed techniques for automatically extracting three different types of metadata from the CMSP documents:

- Publication metadata with titles, authors, and their affiliated institutions from research articles;
- Investigator metadata with name, role, designation, and affiliation of each person from the conference proceedings rosters; and
- Study section metadata with names and affiliations of CMSP program reviewers from separate study section rosters.

We then used the metadata to implement data-analysis functions for discovering patterns and trends in factors such as important drugs, discoveries, investigators, and international collaborations under the CMSP program over its 50-year span.

In FY2013, we explored building a knowledgebase from the metadata of an archived collection and performing semantic queries against such archives. We hoped to discover important domain-specific information, thus generalizing the data-analysis capability. We selected open-source tools to create a model based on the OWL/RDF structural framework for organizing information for a given collection characterized by its metadata fields and their interrelations, and we developed new tools to generate a knowledgebase from the stored metadata using that framework. We used techniques to semantically query such knowledge bases through a Web browser and then to graphically display the results.

We applied this methodology to the CMSP collection by building a complete CMSP knowledgebase and performing semantic queries to obtain various patterns and trends of interest. We also developed an ontology model for the FDA collection that can be used to generate the corresponding knowledgebase once the collection is fully archived.

RIDeM/InfoBot

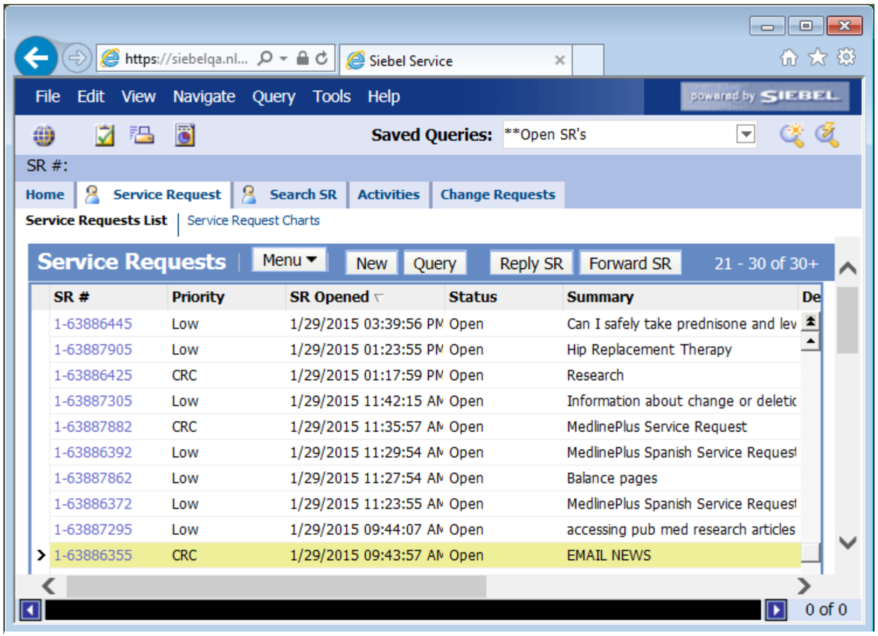
As part of the Clinical Information Systems effort, the RIDeM (Repository for Informed Decision Making) project seeks to automatically find and extract the best current knowledge in scientific publications. The knowledge is provided to several applications (Open-i, a multimodal literature-retrieval engine; Interactive Publications; and InfoBot) through RESTful Web services.

The related InfoBot project enables a clinical institution to automatically augment a patient's electronic medical record (EMR) with pertinent information from NLM and other information resources. The RIDeM API developed for InfoBot allows just-in-time access to patient-specific information to be integrated into an existing EMR system. Such patient-specific information includes medications linked to lists of medications for each patient, or formularies, and images of pills, evidence-based search results for patients' complaints and symptoms, and MedlinePlus information for patient education. For clinical settings without access to the API, a Web-based interface allows information requests to be entered manually.

The InfoBot API integrated with the NIH Clinical Center's EMR system (CRIS) has been in daily use through the Evidence-Based Practice tab in CRIS since July 2009. In 2014, the tab was accessed 615 times a month, on average, by more than 1,380 unique users at the NIH Clinical Center.

Consumer Health Information and Question Answering (CHIQA) System

NLM's customer services receive about 100,000 requests a year. In FY2012, we started to investigate the possibility of automating the process of answering these consumer health questions. In FY2013, we developed and evaluated a prototype CHIQA system. The prototype can classify the incoming requests as either questions about health problems or requests to correct MEDLINE citations. Once the request type is recognized, CHIQA generates an answer and submits it to NLM's reference staff for review. For MEDLINE correction requests, the system automatically finds and retrieves the citation that set off the request, extracts relevant information, and generates an answer. The prototype also understands simple frequently asked questions about causes, treatments, and prognoses of diseases. For these questions, CHIQA finds relevant articles from NLM consumer resources, such as Genetics Home Reference and MedlinePlus, and uses sections of the articles to answer the questions.



SR #	Priority	SR Opened	Status	Summary	De
1-63886445	Low	1/29/2015 03:39:56 PM	Open	Can I safely take prednisone and lev	
1-63887905	Low	1/29/2015 01:23:55 PM	Open	Hip Replacement Therapy	
1-63886425	CRC	1/29/2015 01:17:59 PM	Open	Research	
1-63887305	Low	1/29/2015 11:42:15 AM	Open	Information about change or deletic	
1-63887882	CRC	1/29/2015 11:35:57 AM	Open	MedlinePlus Service Request	
1-63886392	Low	1/29/2015 11:29:54 AM	Open	MedlinePlus Spanish Service Request	
1-63887862	Low	1/29/2015 11:27:54 AM	Open	Balance pages	
1-63886372	Low	1/29/2015 11:23:55 AM	Open	MedlinePlus Spanish Service Request	
1-63887295	Low	1/29/2015 09:44:07 AM	Open	accessing pub med research articles	
1-63886355	CRC	1/29/2015 09:43:57 AM	Open	EMAIL NEWS	

[redacted]

Figure 11. *The Consumer Health Information and Question Answering system has automatically prepared 4,000 responses for review by customer services staff.*

In May 2014, the PubMed Corrections Assistant was integrated into NLM's customer service workflow. It automatically prepares stock replies for NLM customers. Fifty percent of replies generated by the CHIQA PubMed Corrections Assistant do not need editing by customer services staff. Research into improving question classification and recognition is ongoing.

De-identification Tools

De-identification allows the clinical research community to study clinical data without breaching patient privacy. We are developing a clinical text de-identification system called NLM-Scrubber to automatically remove patient identifiers from narrative clinical reports. The provisions of the Privacy Rule of the Health Insurance Portability and

Accountability Act require the removal of 18 individually identifiable information elements that could be used to identify the individual or the individual's relatives, employers, or household members. We completed a version of the software system to be tested at the NIH Clinical Center. This version de-identifies clinical narrative text in a form of electronic messaging known as Health Level Seven (HL7) version 2. It can use information embedded in various HL7 fields as well as externally provided information, such as the list of names of the healthcare providers at NIH.

We are using another tool we developed – the Visual Tagging Tool (VTT) – to develop a collection that we are using as the gold standard for developing and testing NLM-Scrubber. By the end of 2014, we amassed a collection of 23,446 clinical reports of 8,150 patients, in which human reviewers manually labeled every piece of individually identifiable information. We are working on deploying the NLM-Scrubber to the NIH Clinical Research Information System (CRIS), which is NIH's clinical research repository. All clinical narrative reports generated by NIH clinicians and researchers will be de-identified using NLM-Scrubber. We also shared the VTT with the natural language processing community to be used for other types of lexical tagging and text annotation.

Librarian Infobutton Tailoring Environment (LITE)

Infobuttons are links from one information system to another that anticipate users' information needs, take them to appropriate resources, and help them retrieve relevant information (http://www.infobuttons.org). They are mostly found in clinical information systems (such as electronic health records (EHRs) and personal health records (PHRs)) to give clinicians and patients access to literature and other resources that are relevant to the clinical data they are viewing. The NIH Clinical Center Laboratory for Clinical Informatics Development has worked with Health Level Seven, Inc. (HL7) — an electronic messaging standards-development organization — to develop an international standard to support the communication between clinical systems and knowledge resources. MedlinePlus Connect currently provides an HL7-compliant query capability.

To increase the usefulness of infobuttons, they are typically linked not to a specific resource, but instead to an "infobutton manager" that uses contextual information (such as the age and gender of the patient, the role of the user, and the clinical data being reviewed) to select the most applicable resources from a large library of known resources. The infobutton manager customizes the links to those resources using appropriate data from the context, and then presents the user with a list of those links. The NIH Clinical Center Laboratory for Informatics Development is working with investigators at the University of Utah and the Department of Veterans Affairs to establish a freely available, HL7-compliant infobutton manager, known as "Open Infobutton," that can be a national resource for electronic health record developers and users (http://www.openinfobutton.org). With the Open Infobutton, clinicians and patients will be able to obtain the health-related information they need, when and where they need it.

Infobutton managers, including Open Infobutton, require knowledge bases to do their customization work. The knowledge bases are very institution-specific. We developed the LITE, a user-friendly tool that an institution's medical librarians can use and that provides Open Infobutton with the knowledge it needs to customize its responses to requests from that institution. The system is in beta testing at the University of Utah (http://lite.bmi.utah.edu). In 2013, we transferred the maintenance of LITE to the University of Utah, which will continue to make it open-access and will also develop the software to make it part of an open-source package that can be installed at any institution. A user-evaluation project is now under way in collaboration with Ohio University.

Terminology Research and Services

The Patient Data Management Project (PDM) brings together several activities centered on lexical issues, including developing and maintaining the SPECIALIST lexicon and lexical research. Those lexicon and lexical tools support key NLM applications. A package of lexical-tool applications underlies the MetaMap algorithm we use to find UMLS concepts in biomedical text and to automatically index MEDLINE abstracts. We distribute the lexicon and lexical tools to the medical informatics community as free open-source tools.

The 2014 release of the SPECIALIST lexicon contains 484,628 records representing more than 896,000 forms, an increase of 7,771 records from the 2013 release. Many of the new records are derived from de-identified clinical records from our own de-identification project and from our work with the MIMIC-II database. Others were identified using an n-gram database of MEDLINE.

Semantic Knowledge Representation

The Semantic Knowledge Representation (SKR) project conducts basic research, based on the UMLS knowledge sources, in symbolic natural language processing. A core resource is the SemRep program, which extracts semantic predications (relationships — such as interacts with, treats, causes, inhibits, and stimulates — between drug and disease, gene and gene, gene and disease, drug and drug, etc.) from biomedical text. SemRep was originally developed for biomedical research. We're developing a way to extend its domain to influenza epidemic preparedness, health promotion, and health effects of climate change. In FY2013, we made a downloadable version of SemRep available to the public.

SemRep finds biomedical-related semantic relationships in MEDLINE, and then our Semantic MEDLINE Web application manipulates those relationships. The SKR project maintains a database of more than 70 million SemRep predications extracted from all MEDLINE citations; the database is available to the research community. This database supports Semantic MEDLINE, which integrates PubMed searching, SemRep predications, automatic summarization, and data visualization. The application helps users manage the results of PubMed searches by creating an informative graph with links to the original MEDLINE citations and by providing convenient access to additional relevant knowledge resources (such as Entrez Gene, the Genetics Home Reference, and the UMLS® Metathesaurus®). The Semantic MEDLINE technology was recently adapted for analyzing NIH grant applications, allowing NIH portfolio analysts to track emerging biomedical research trends and identify innovative research opportunities.

Clinical Vocabulary Standards and Associated Tools

Many of our projects in this area continue to promote the development, enhancement, and adoption of clinical vocabulary standards. During FY2014, we performed extensive work related to NLM-developed and -supported clinical vocabulary and messaging standards that are crucial to the success of the Federal Health Information Technology (HIT) goals. We provided comments and assistance to the Office of the National Coordinator for HIT (ONC), Centers for Medicare and Medicaid Services (CMS), and FDA, often on extremely short deadlines, about meaningful-use regulations for electronic health record vocabulary and clinical-data-standards requirements, with special focus on improving the fit of quality measures to the realities of EHRs and avoiding efficiency losses in clinical office practices. LHNCBC's Director served on the Trans-NIH Outcomes-Effectiveness Research Interest Group, the National Center for Advancing Translational Sciences (NCATS) Global Rare Disease Registry Steering Committee and Common Data Elements Working Group, the National Children's Study Information Management System Review, and the American Medical Informatics Association (AMIA) EHR-2020 Task Force. He was also selected to serve on the Content Standards Workgroup of the HIT Standards Committee — the Federal Advisory Committee charged with making recommendations to ONC on standards, implementation specifications, and certification criteria for the electronic exchange and use of health information.

In FY2014, we participated in an effort to align SMART, CIMI, and FHIR behind NLM-developed and -supported clinical vocabularies LOINC (Logical Observation Identifiers, Names, and Codes – for medical tests, measurements, and observations), SNOMED® (Systematized Nomenclature of Medicine — Clinical Terms), RxNorm (for medications), and UCUM (for computable units of measure). Our SNOMED CT® mapping projects, described below, support Federal Meaningful Use Stage 2 (MU2) EHR requirements to use SNOMED CT for coding diagnoses.

Medical Ontology Research

The Medical Ontology Research (MOR) project focuses on basic research on biomedical terminologies and ontologies and their application to natural language processing, clinical decision support, translational medicine, data integration, and interoperability.

During FY2014, we investigated the coverage and representation of human phenotypes in standard terminologies and developed unsupervised machine-learning methods for extracting drug-drug interactions from the structured product labels. We also developed a prototype version of MeSH in RDF (Resource Description Framework) for use in the Semantic Web as a Linked Open Data resource. We continued supporting the NLM Value Set Authority Center, focusing on the quality of drug value sets in clinical quality measures. Finally, in collaboration with the Center for Drug Evaluation and Research at the Food and Drug Administration (FDA), we developed methods for extracting adverse drug events from MEDLINE indexing.

The CORE Problem List Subset of SNOMED CT

SNOMED CT is a comprehensive, multilingual medical terminology for anatomic sites, organisms, chemicals, diagnoses, symptoms, findings, and other such concepts. The problem list — a patient's list of active conditions and symptoms — is an essential part of the electronic health record (EHR). Meaningful-use regulations from CMS require the use of SNOMED CT to code the problem list and many other EHR fields.

We analyzed problem-list vocabularies and their usage frequencies in seven large-scale U.S. and overseas healthcare institutions, identified a subset of the most frequently used problem-list terms in SNOMED CT, and then published it as the CORE (Clinical Observations Recording and Encoding) Problem List Subset of SNOMED CT. The CORE Subset can be a starter set for institutions that do not yet have a problem-list vocabulary, and this will save significant development effort and reduce variations between institutions. Existing problem-list vocabularies can also be mapped to the CORE Subset to facilitate data interoperability.

Since its first publication in 2009, the CORE Subset has received considerable attention from the IHTSDO (International Health Terminology Standards Development Organization), the SNOMED CT user community, EHR software vendors, and terminology researchers. It has been installed in various EHR products and used as a focus for SNOMED CT-related research, mapping projects, and quality assurance. The MedlinePlus Connect Project, which facilitates linkage by medical records systems and other outside sources to NLM's rich consumer sources of medical information, has mapped the CORE Subset to MedlinePlus health topics so that medical records systems could automatically pull this educational material for patients to use. In 2012, the CORE Subset was enriched with a clinical dataset from the U.S. Department of Veterans Affairs covering 33 million patients.

In FY2014, we maintained and updated the CORE Problem List Subset of SNOMED CT, which now contains 6,100 concepts and has been used in EHRs, terminology research, and quality assurance activities. In addition to the standalone downloadable file, the CORE Problem List Subset is also available as a reference set (refset) in the international release of SNOMED CT and in the UMLS as a specific content view. It is updated four times a year to synchronize with changes in SNOMED CT and the UMLS.

Mapping Between SNOMED CT and ICD Codes

International Classification of Diseases (ICD) codes are required for public health reporting of population morbidity and mortality statistics. In the United States, ICD-9-*CM* (the ninth version of "Clinical Modification") is also used for reimbursement (replaced by ICD-10-*CM* in October 2014). Because of this need, many existing EHR systems are still using ICD-based vocabularies to encode clinical data. However, ICD was not designed to capture information that is detailed enough to support clinical care. SNOMED CT is a much better clinical terminology for that purpose, and its use will be required as part of the meaningful-use regulations.

To encourage the migration to SNOMED CT and to enable EHRs to output ICD codes for administrative purposes, we have developed various maps between SNOMED CT and the ICD classifications. We published a SNOMED CT-to-ICD-10-*CM* rule-based map that allows users to encode patient problems in SNOMED CT terms and then generate the appropriate ICD-10-*CM* codes in real time for billing or other purposes. In FY2014, we expanded the coverage of the SCT-to-ICD-10-*CM* map to 54,000 concepts (from 35,000 in FY2013). The map has been used by 1,200 UMLS licensees. To demonstrate the use of the map, we developed the I-MAGIC (Interactive Map-Assisted Generation of ICD Codes) demo tool. In FY2014, I-Magic attracted 22,281 visitors, up from 15,271 in FY2013.

For an international project, in collaboration with the IHTSDO and the World Health Organization (WHO), we helped develop an analogous rule-based map from SNOMED CT to ICD-10. We also adapted our I-MAGIC tool to showcase this map. In a separate project, to help convert legacy ICD-9-*CM*-encoded clinical data into SNOMED CT codes, we produced two more maps to SNOMED CT, from 9,000 commonly used ICD-9-*CM* diagnostic codes and from 3,000 ICD-9-*CM* procedure codes.

We are currently investigating the need to create maps between SNOMED and ICD-10-*PCS* (Procedure Coding System), since SNOMED CT is also designated as the terminology standard for coding clinical procedures in Phase 2 of the meaningful-use incentive program for electronic health records.

RxTerms

RxTerms is a free, user-friendly, efficient interface terminology for drugs that links directly to RxNorm, the national terminology standard for clinical drugs. The Centers for Medicare and Medicaid Services (CMS) used RxTerms in one of their pilot projects in the post-acute care environment. RxTerms is also used in the NLM PHR, at least one

EHR from a major medical institution in Boston, and by some proprietary-drug-software vendors. During FY2012, we aligned the data model of RxTerms and RxNorm by creating a new term type in RxNorm to cover the drug-route combination. We're continuing to align data elements between RxTerms and RxNorm, and we're reviewing ways to improve the display names of multi-ingredient generic drugs to improve usability. We update RxTerms monthly to synchronize with the full release of RxNorm. In addition to the standalone download, RxTerms is available through the RxNorm API.

RxNav

Released in September 2004, RxNav was first developed as a graphical interface browser to the RxNorm database and was primarily designed for displaying relations among drug entities. In addition to the RxNav browser, we later created SOAP-based and RESTful application programming interfaces (APIs) to let users integrate RxNorm functions into their applications. We also integrated additional drug information sources, including RxTerms, the National Drug File-Reference Terminology (NDF-RT), and pill images.

During FY2014, we developed RxClass-an application (and companion API) for linking RxNorm drugs to drug classes from the Anatomical Therapeutic Chemical (ATC) classification of drugs-from the Medical Subject Headings (MeSH) and DailyMed. We also investigated drug-drug interaction information from DrugBank instead of from NDF-RT, where this information is no longer available. RxNav, RxClass, and the APIs received a combined total of about 230 million queries during FY2014 (a significant increase from 70 million queries in FY2013). Users include clinical and academic institutions, as well as pharmacy management companies, health insurance companies, EHR vendors, and drug information providers. Developers of mobile apps have also started to integrate our APIs into their applications.

LOINC® Standards for Identifying Clinical Observations and Orders

Within medical record systems, patient summaries, and reports to public health organizations, Federal Meaningful Use Stage 2 (MU2) EHR regulations require that lab result messages sent to ordering clinicians use LOINC (Logical Observation Identifiers, Names, and Codes). In FY2014, we continued to work with the Regenstrief Institute, major laboratory companies, several NIH Institutes, and other organizations to expand the size and breadth of the LOINC database. The FDA Mini-Sentinel program is adopting LOINC for its lab codes, and the National Patient-Centered Clinical Research Network (PCORnet) is encouraging the use of LOINC for all variables. A hospital in Buenos Aires is now using LOINC, Infobutton, and NLM's MedlinePlus Connect to deliver information about lab tests to patients' personal health records.

By the end of FY2014, LOINC had more than 32,000 users in 162 countries and had been translated into 15 languages, including new translations into Dutch, French, and Portuguese. Users of the LOINC user browsing tool can pick any of these languages, search for words in the chosen language, and see the matching LOINC terms in that language plus English. The Regenstrief Institute and the International Health Terminology Standards Development Organisation (IHTSDO) signed a long-term agreement to begin collaborative work on linking their leading global healthcare terminologies: LOINC and SNOMED Clinical Terms.

We worked with Regenstrief and the LOINC Committee to create more than 1,200 new LOINC terms for both laboratory and clinical variables, and the LOINC database now contains nearly 74,000 terms. We released new terms for toxicology, chemistry, hematology, microbiology, and clinical survey assessment instruments (e.g., AHRQ's (Agency for Healthcare Research and Quality's) Healthcare Event Reporting Form).

We continued to meet with other NIH organizations that are developing assessment instruments and registry system values, with the goal of closer alignment among NIH standard element development efforts. We are collaborating with other NIH organizations (and the Regenstrief Institute) to structure their assessment instruments and registry system values into the LOINC format and incorporate them into the LOINC database, a common framework that includes many kinds of clinical and research variables. We serve on the Common Data Elements (CDE) Working Group to the trans-NIH BioMedical Informatics Coordinating (BMIC) Committee. We're also working with colleagues at these institutes:

- National Eye Institute (NEI), to restructure its packages of assessment instruments for the National Ophthalmic Disease Genotyping Network (eyeGENE®);
- National Center for Advancing Translational Sciences (NIH/NCATS) Office of Rare Diseases Research, to revise the CDEs for their registry system for rare diseases — and we plan to create corresponding LOINC codes;

- National Heart, Lung, and Blood Institute (NHLBI), NICHD, and NIH/NCATS, on the NHLBI Hemoglobinopathies Uniform Medical Language Ontology (HUMLO) project, to develop CDEs for hemoglobinopathies using standard terminologies such as LOINC for the questions and SNOMED CT for the answers; and
- National Institute of Neurological Diseases and Stroke (NINDS), on CDEs and LOINC codes for the neurological quality-of-life (Neuro-QOL) measures done a few years ago.

Next-Generation Electronic Health Records to Facilitate Patient-Centric Care, Clinical Research, and Public Health

These projects target the overall recommendations of the NLM Long Range Plan Goal 3: Create Integrated Biomedical, Clinical, and Public Health Information Systems that Promote Scientific Discovery and Speed the Translation of Research into Practice.

Big Data to Knowledge (BD2K): Using Large Clinical Databases to Assess the Associations among Patient Factors, Medication Usage, and Patient Outcomes

These projects support NIH's Big Data to Knowledge (BD2K) initiative and the NIH Director's research priorities. We used large clinical databases to predict patient outcomes from patient factors, including medication usage. We combined structured data and information-retrieval techniques and developed and implemented algorithms. Ongoing studies illustrate the potential for Big Data to generate knowledge vital to improving clinical care.

Getting an accurate medication history for Emergency Department (ED) patients is important for their emergency care, especially since a significant proportion of ED visits are related to adverse events from prescription medications. Gathering such information from patients is time-consuming, expensive, and sometimes impossible (such as when a patient is unconscious), and patient-provided medication histories are often incomplete. The first phase of this study was to assess the degree to which data from a national prescription database enhanced the routine medication history and the percent of ED patients who had such data. Our results showed that the national prescription database (which could be used widely) increased the completeness of patients' medication history by 28 percent compared with the manual history alone, but such information was available for only 60 percent of the ED patients. These results were published in September 2013.

We have been working with MIMIC II for a number of years. MIMIC II is a database organized by MIT and funded by NIBIB and NIGMS that collects almost all medical record data from ICU admissions at one large hospital, de-identifies the data, and organizes it in a usable database. It includes all ICU nursing variables vital signs, orders, and information about local interventions, medications prescribed and dispensed, nursing notes, and radiology reports. It has physician admission and progress notes for the NICU babies, and otherwise only discharge summaries. It also includes all laboratory results produced during the hospital stay at which the ICU admission occurred. The current version of MIMIC II carries nearly 280 million rows of data about 32,536 patients (8,087 are neonates and the rest adults), plus waveform files. We are expecting a new version from MIT that will add about 50% more patients and data. In the earliest versions this data was provided as local codes that were very duplicative. For example, there were more than 30 distinct codes for serum glucose. Here at LHNCBC, we translated all of their many different local codes to universal codes—LOINC for lab test and clinical measures, RxNorm for drugs and SNOMED CT for problems. We returned this data to MIT and they incorporated it into their distributed version so that all of the MIMC II users could take advantage of these codings. MIMIC II also carries some information collected from outside of the ICU, including vital status (alive or dead) obtained by the originating hospital before de-identification via linkage to the Social Security death tapes.

We have completed a number of studies on the MIMIC II data (PMIDs: 24384230, 24551317, 23613065, 23249446). Investigators are now looking at the relation between use of statin medications by ICU patients and the effect on mortality during sepsis, because the anti-inflammatory effects of statins are postulated to be protective in such situations. One of our fellows looked at the concordance between a well-known formula based on laboratory test for predicting Disseminated Intravascular Coagulation (DIC) in ICU patient and the diagnoses and therapeutic actions taken by providers. This work is presented at the 2015 Joint Summit of the American Medical Informatics Association. Using neonatal intensive care unit data for 75 infants with necrotizing enterocolitis (NEC) and 189 controls thoroughly matched on gestational age, gender and being small for gestational age, we have established the role of feeding in NEC. Addition of cow milk-based products to breast milk significantly increased the risk of NEC for premature infants. Infants receiving combination feeds had more than triple the risk of developing NEC as those on exclusive breast milk feeds, and those receiving only formula had seven times the risk.

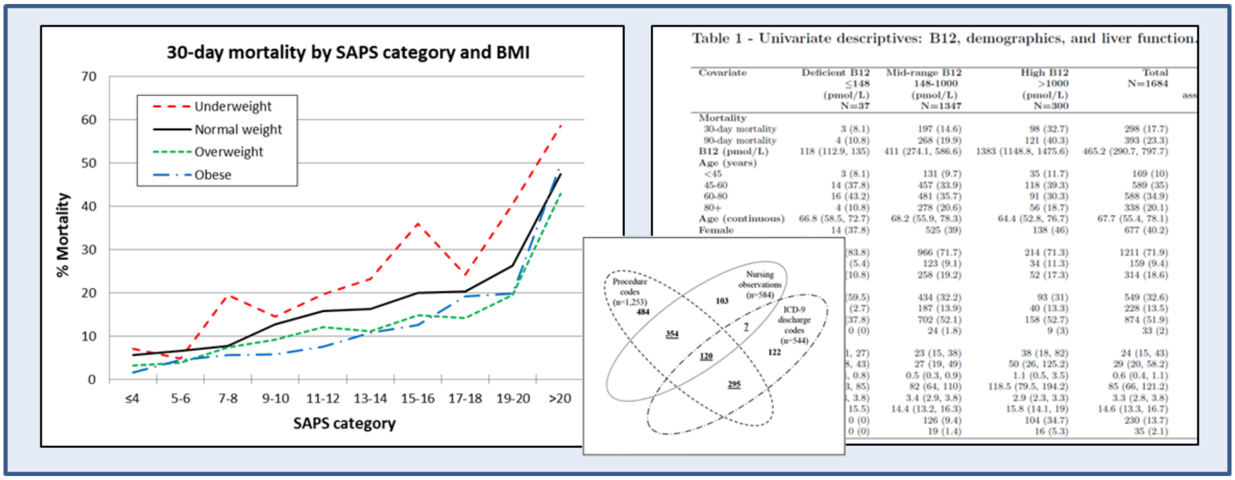


Table 1 - Univariate descriptives: B12, demographics, and liver function.

Covariate	Deficient B12 ≤148 (pmol/L) N=37	Mid-range B12 148-1000 (pmol/L) N=1347	High B12 >1000 (pmol/L) N=300	Total N=1684
Mortality				
30-day mortality	3 (8.1)	197 (14.6)	98 (32.7)	298 (17.7)
90-day mortality	4 (10.8)	268 (19.9)	121 (40.3)	393 (23.3)
B12 (pmol/L)	118 (112.9, 135)	411 (274.1, 586.6)	1383 (1148.8, 1475.6)	465.2 (290.7, 797.7)
Age (years)				
<45	3 (8.1)	131 (9.7)	35 (11.7)	169 (10)
45-60	14 (37.8)	457 (33.9)	118 (39.3)	589 (35)
60-80	16 (43.2)	481 (35.7)	91 (30.3)	588 (34.9)
80+	4 (10.8)	278 (20.6)	56 (18.7)	338 (20.1)
Age (continuous)	66.8 (58.5, 72.7)	68.2 (55.9, 78.3)	64.4 (52.8, 76.7)	67.7 (55.4, 78.1)
Female	14 (37.8)	525 (39)	138 (46)	677 (40.2)
	(83.8)	966 (71.7)	214 (71.3)	1211 (71.9)
	(5.4)	123 (9.1)	34 (11.3)	159 (9.4)
	(10.8)	258 (19.2)	52 (17.3)	314 (18.6)
	(59.5)	434 (32.2)	93 (31)	549 (32.6)
	(2.7)	187 (13.9)	40 (13.3)	228 (13.5)
	(37.8)	702 (52.1)	158 (52.7)	874 (51.9)
	0 (0)	24 (1.8)	9 (3)	33 (2)
	1, 27)	23 (15, 38)	38 (18, 82)	24 (15, 43)
	8, 43)	27 (19, 49)	50 (26, 125.2)	29 (20, 58.2)
	, 0.8)	0.5 (0.3, 0.9)	1.1 (0.5, 3.5)	0.6 (0.4, 1.1)
	3, 85)	82 (64, 110)	118.5 (79.5, 194.2)	85 (66, 121.2)
	, 3.8)	3.4 (2.9, 3.8)	2.9 (2.3, 3.3)	3.3 (2.8, 3.8)
	15.5)	14.4 (13.2, 16.3)	15.8 (14.1, 19)	14.6 (13.3, 16.7)
	0 (0)	126 (9.4)	104 (34.7)	230 (13.7)
	0 (0)	19 (1.4)	16 (5.3)	35 (2.1)

Figure 12. *In 2014, we continued researching secondary use of large collections of clinical data from the MIMIC II Intensive Care Unit database.*

We obtained access to the CMS Enclave that provides analytic (but not direct) access to the new and improved version of Medicare and Medicaid data and patient assessments, called "CMS Chronic Conditions Data Warehouse (CCW)." The CCW is a research database designed to make Medicare, Medicaid, Assessments, and Part D Prescription Drug Event data more readily available to answer research questions. The beauty of the CCW compared to the previous state is that beneficiary matching, deduplication, and merging of the files in preparation for a study – which in the past had to be done by each researcher – has already been done by CMS. Such data can be directly downloaded to give researchers patient level de-identified data at their site, but direct downloading can be expensive, and the IRB permissions can be difficult to obtain. With the Enclave, researchers can analyze the CCW data while the data remains in place on CMS's computer, via SQL queries and SAS, but cannot see (or download) any patient level data or aggregations of very small samples. Because of this extra layer of privacy protection, permissions are easier to get, and for other reasons the cost is much less. The CCW Beneficiary identifier field is a unique key specific to the CCW and is not applicable to any other identification system or data source, which allows researchers to link and analyze information across the continuum of care. CCW also more rapidly updates than the old system, with most files being up-to-date 1-2 weeks after a visit encounter. It contains 100% of Medicare files from 1999 to 2013, and comparable Medicaid data but only from 2011. This is a truly humongous database with more than 12 billion individual records. We obtained permission to analyze a large subset of the database with a special focus on drugs, treatments, and Part D. It took some time to get connected and find our way around the very large and complicated database. We have begun projects looking at the use of influenza vaccine based on charges for the dispensing, the effect of better adherence to the current recommendations for when to get the flu shot (considering the actual peaks of incidence), and at the overall rate of flu shots dispensed compared to patients reports of usage in the Medicare survey.

Newborn Screening Coding and Terminology Guide

We've collaborated with many federal, state, and other agencies to standardize the variables used in newborn screening (NBS) by using national coding standards as required by Meaningful Use Stage 2 (MU2). Our collaborators include the Health Resources and Services Administration (HRSA), the Centers for Disease Control and Prevention (CDC), the Association of Public Health Laboratories (APHL), the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), and NHLBI. We created a comprehensive panel of LOINC terms for NBS and continue to create new LOINC terms as new conditions and tests come into play. We also periodically review and update existing codes based on user feedback.

During FY2014, we worked closely with many states and vendors on their implementation of newborn screening LOINC and SNOMED CT coding and HL7 messaging standards. We revised many LOINC terms and created several new terms based on requests from NBS programs and laboratories. We created a comprehensive LOINC panel to report screening results for critical congenital heart disease (CCHD), the latest condition added to the HHS Secretary's Recommended Uniform Screening Panel for NBS. We expanded beyond NBS and worked with many states to create a panel of terms for reporting the results of therapeutic diet monitoring for patients with conditions that were diagnosed based on NBS. We also participated as a technical advisor about terminology and interoperability standards for HRSA's CCHD-pilot-program grantees and represent NLM on the Newborn Screening Technical assistance and Evaluation Program (NewSTEPs) Steering Committee and Health Information Technology workgroup.

Information Resource Delivery for Researchers, Care Providers, and the Public

We perform extensive research on developing advanced computer technologies to facilitate the access, storage, and retrieval of biomedical and consumer health information.

ClinicalTrials.gov

Established in 2000, ClinicalTrials.gov makes comprehensive information about registered clinical research studies readily available to the public. Each month during the past year, it received more than 135 million page views and hosted about 960,000 unique visitors. Nearly 16,200 study sponsors, including the federal government, pharmaceutical and device companies, and academic and international organizations, submitted data to ClinicalTrials.gov through the Web-based Protocol Registration and Results System (PRS). At the end of FY2014, the site had more than 176,200 research studies, conducted in all 50 states and in more than 180 countries. Approximately one-third of the studies were still open to recruitment. For the remaining two-thirds, the recruitment phase was either over or the study had been completed. Summary-results tables describing primary and secondary outcomes, adverse events, and characteristics of the participants were posted for more than 14,600 of the registered studies.

In FY2014, new registrations of clinical trials were submitted at an average rate of 460 records per week, an increase of 15 percent from FY2013. The average rate of new results submissions was about 90 per week, an increase of 29 percent from FY2013. We attribute the continued growth in the use of ClinicalTrials.gov to U.S. laws that require registering and reporting the summary results of clinical trials, as well as international recognition of the scientific and ethical importance of registering and reporting results. The combined registry and results database provides information about ongoing and completed clinical research for patients, healthcare providers, and policy decision makers.

ClinicalTrials.gov staff continued to educate the public about the most recent federal law concerning clinical trials and results submission, Section 801 of the Food and Drug Administration Amendments Act of 2007 (FDAAA 801). We continued to work with the NIH Office of the Director, other NIH Institutes and Centers, and the FDA on a Notice of Proposed Rulemaking (NPRM) that elucidates the requirements of FDAAA 801 as well as key implementation issues. We also participated in the development of an NIH draft policy on the registration and results submission of all NIH-funded clinical trials. (Both the NPRM and the NIH draft policy were issued in November 2014 for public comment.)

Since FY2013, we have been evaluating and enhancing the PRS site. Sponsors and investigators use the site to submit, update, and maintain study registration and summary-results information. On the basis of direct user feedback, findings from usability studies, and recommendations from usability experts, we continued implementing user-interface improvements intended to streamline the data-entry process in FY2014. We also continued providing targeted education and outreach on the results database and submission requirements through developing additional help materials, presenting at conferences, participating in working groups, and publishing in journals.

ClinicalTrials.gov research projects and publications in FY2014 included:

- Identifying reporting discrepancies between results of trials posted on ClinicalTrials.gov and corresponding findings published in the peer-reviewed literature.
- Using ClinicalTrials.gov data to estimate the "gap" in U.S. federal human research oversight regulations.

Characterizing prematurely terminated trials registered on ClinicalTrials.gov and determining whether primary outcome data are available from such trials in the ClinicalTrials.gov results database and the published literature.

Genetics Home Reference (GHR)

The GHR Web site offers high-quality information about genetic conditions and the genes and chromosomes related to those conditions. It answers the public's questions about human genetics by using the rich technical data from the Human Genome Project and other genomic research. At the end of FY2014, GHR offered user-friendly summaries of 2,376 genetics topics, including more than 1,000 genetic conditions, 1,240 genes, all the human chromosomes, and mitochondrial DNA. GHR also offers an online handbook called Help Me Understand Genetics, which provides an illustrated introduction to fundamental topics in human genetics, including mutations, inheritance, genetic testing, gene therapy, and genomic research.

In the past year, the GHR team expanded the Web site's genetics content for consumers, adding 252 new genetics summaries and two new Help Me Understand Genetics pages about genetic susceptibility and informed consent. We also reviewed and updated 157 existing summaries. The team hosted a "virtual" intern from the University of Maryland, who helped us add and update links to consumer resources throughout the Help Me Understand Genetics handbook. In FY2014, the GHR Web site averaged 16.5 million page views per month and 61,500 visitors per day (an increase of 22 percent and 45 percent, respectively, from FY2013).

In support of SNOMED CT's initiative to include more rare and genetic diseases, we submitted more than a dozen updates for existing terms and synonymy in SNOMED CT (the Systematized Nomenclature of Medicine — Clinical Terms). We also continued to integrate GHR results into NLM's MedlinePlus Connect. This service enables electronic medical records and other applications that use MedlinePlus Connect to retrieve GHR summaries (along with MedlinePlus content) by using code queries from SNOMED CT and, beginning in April 2014, from ICD-10-CM (International Classification of Diseases, 10th Revision, Clinical Modification). GHR topics, each of which can map to multiple SNOMED CT codes, mapped to more than 2,800 SNOMED CT codes at the end of FY2014.

This year, GHR began a new contract with Genetic Alliance, an umbrella organization for condition-specific genetics interest groups. The new contract continues our three-year collaboration with Genetic Alliance to update existing GHR Web site content and track new research developments about particular genetic conditions. We updated about 60 existing GHR topics through this initiative in FY2014. In another collaboration, the GHR team consulted with researchers at the National Human Genome Research Institute (NHGRI) to develop content for use in a study involving genetic counseling for carriers of autosomal recessive genetic diseases.

Beginning in July 2014, GHR implemented a customer satisfaction survey in partnership with ForeSee, Inc. The initial survey, which runs for one year, will provide information that the team can use to improve the content, look, and navigation of the GHR Web site. We also added "AddThis" buttons to each page of the Web site this year, allowing users to share GHR content via e-mail, Facebook, Twitter, and other social media. Additionally, we performed outreach activities to increase public awareness about GHR. For example, we presented the Web site to several groups, including health and science journalists who visited NLM as part of the Association of Health Care Journalists–NLM Fellowship program and clinical and molecular fellows at NHGRI.

Profiles in Science Digital Library

The Profiles in Science® Web site showcases digital reproductions of items selected from:

- The personal manuscript collections of five prominent biomedical researchers, doctors, public health practitioners, philanthropists, political leaders, and other people who provided resources, removed barriers, and spearheaded projects to improve the health of the nation and the world and
- Three thematic collections: the 1964–2000 Reports of the Surgeon General, the history of the Regional Medical Programs, and the Visual Culture and Health Posters.

The site gives researchers, educators, and future scientists all over the world access to unique biomedical information previously accessible only by making in-person visits to the institutions holding the physical manuscript collections. It also serves as a tool for recruiting donations of collections from scientists who wish to preserve their papers for future generations. It decreases the need for handling the original materials by making available high-quality digital surrogates of the items. Standardized, in-depth descriptions of each item make the materials widely accessible, including to individuals with disabilities. The growing Profiles in Science digital library provides ongoing opportunities for future experimentation in digitization, optical character and handwriting recognition, automated image identification, item description, digital preservation, emerging standards, digital library tools, and search and retrieval.

The content of Profiles in Science is created in collaboration with the History of Medicine Division of NLM, which processes and stores the physical collections. Collections donated to NLM contain published and unpublished materials, including manuscripts, diaries, laboratory notebooks, correspondence, photographs, poems, drawings, and audiovisual resources.

In FY2014, we added two new collections. Sir William Osler (1849–1919) revolutionized American medical education with the clinical internship program he instituted at the new Johns Hopkins School of Medicine. Senior students worked in hospital wards full time to "learn medicine at the bedside." The Osler addition was made possible through collaboration with McGill University and Johns Hopkins University. Mike Gorman (1913–1989) was a well-known journalist, author, publicist, and crusader for health policy reform who won a Lasker Award in 1948 for his newspaper exposés of state mental hospital conditions in Oklahoma.

The Web site averages more than 105,000 unique visitors each month, including people seeking an authoritative source of information about current events. The Web site experienced a spike in interest during the HHS and White House celebrations of the 50th anniversary of the Surgeon General's 1964 Report on Smoking and Health.

At the end of FY2014, the 388 publicly available collections contained 27,531 items composed of 144,594 digitized image pages, including transcripts of 11,612 handwritten pages or pages we couldn't use optical character recognition technology for.

Figure 13. *Screenshots from the biographical video highlighting the life and career of Dr. Michael E. DaBakey for NLM's Profiles in Science program.*

Turning the Pages (TTP)

The goal of this project is to give laypeople access to historically significant and previously inaccessible books in medicine and the life sciences. We build 3D models for books and develop animation techniques that let users touch and turn page images in a realistic way on touch-sensitive monitors in kiosks at NLM or tablets using a high-resolution ("Retina") iPad app, or click and turn for online versions. We've also built a different 3D model for a "scroll"-type document and applied it to the 1700 BC Edwin Smith medical papyrus, which can be "touched" (or clicked) and "rolled out." The TTP Web site is very popular, attracting 10,000 unique visitors a month and 326,000 page views a month. The iPad app is also popular: in FY2014, 4,536 people downloaded books with the app for the first time (new users), and more than 77,130 people who already had the app updated it with the newer version of the software, which indicates a steady user base.

The Turning the Pages kiosks at NLM and the NLM Web site now present 122 rare books. In FY2014, we added two books to the iPad version and one book to the kiosk and Web versions.

For the longer term, we're studying ways to develop a reactive 3D implementation system for TTP and investigating tools for this purpose, such as Unity, Coco's 3D, and the Unreal Engine. The advantages of a real-time 3D system are:

- We can produce 3D versions of each book more quickly,
- Other institutions can use our software to create their own interactive books, and
- We can discover new functionalities, such as rotating a book 360° and turning multiple pages at once.

In addition, anticipating the next generation of kiosk design, we're investigating newer display technologies such as multi-touch monitors, which enable users to use two or more fingers to perform tasks such as zooming in and out.

Evidence-Based Medicine: PubMed for Handhelds (PubMed4Hh)

Developed and released in FY2003, PubMed for Handhelds (PubMed4Hh) facilitates evidence-based medical practice with MEDLINE access from almost anywhere via smartphones, wireless tablet devices, netbooks, and portable laptops. PubMed4Hh requires no proprietary software and reformats the screen display as appropriate for the wireless handheld device being used. Clinical filters feature easy access to relevant clinical literature.

Newly developed resources allow people to search MEDLINE through text messaging. An algorithm to derive "the bottom line" (TBL) of published abstracts allows clinicians to quickly read summaries from almost anywhere. For example, it enables doctors to quickly consult research findings to help determine the best course of treatment for a patient. A "consensus abstracts" element provides rapid review of multiple publications with smartphones. A recent review of PubMed4Hh server logs showed that more than 90 percent of queries were clinical in nature.

To evaluate the usefulness of abstracts in clinical decision making, randomized controlled trials using simulated clinical cases were conducted by the Uniformed Services University of the Health Sciences, the Botswana–University of Pennsylvania partnership, and the National Telehealth Center and Philippine General Hospital, Manila. These studies demonstrated the usefulness of the app for clinical decision making.

The PubMed4Hh app is available for iOS (iPhone/iPad) and Android users. In FY2014, the iOS app was downloaded 240,000 times by users in the United States (45 percent of downloads) and elsewhere (55 percent). Queries from smartphone apps now account for 60 percent of all queries. The total number of searches has tripled since before the smartphone app was introduced.

Interactive Publications

Recognizing the increasing use of multimedia in scientific work, LHNCBC researchers investigate and develop models for highly interactive multimedia documents that could transform the next generation of publishing in biomedicine. The Interactive Publications project focuses on the standards, formats, and authoring and reading tools necessary for creating and using interactive publications that, in addition to text, contain media objects relevant to biomedical research and clinical practice. These objects include video, audio, bitmapped images, interactive tables and graphs, and clinical DICOM (Digital Imaging and Communications in Medicine) image formats for X-rays, CT scans, MRIs, and ultrasounds.

We've developed interactive publications containing these data types and tools for viewing, analyzing, and writing them. The tools, Panorama and Forge, are analogous to Adobe's Acrobat Reader and Professional for PDF documents. Panorama, used for viewing and analyzing these publications, was 1 of 9 semifinalists out of 70 entrants in Elsevier's Grand Challenge contest four years ago. We conducted a formal usability study of Panorama in 2013 and, as a result, enhanced the software to include bar charts and the ability to run on the Mac OS X operating system.

We extended Panorama to provide Annotation Concepts. Clicking on text in an interactive publication prompts an NLM servlet (RIDeM, for Repository for Informed Decision Making, developed in-house in 2004) to identify the corresponding Unified Medical Language System, or UMLS®, concepts. The servlet returns an XML (eXtensible Markup Language) file to Panorama, which parses it to provide the preferred UMLS term and semantic group, and it provides linkouts for MedlinePlus, eMedline, Family Doctor, and other resources. We are continuing to develop additional features to group concepts by semantic relationships and other factors.

Our initial Panorama software was a desktop product requiring a large download, which was an inconvenience for many users. Consequently, we investigated several Web-based methods for reading interactive publications. In 2012, we developed our first browser-based version (Panorama Lite) using Adobe Flex, thus eliminating the need to download the Panorama software. The only requirement to run it is to have Flash installed. In 2013, we improved and updated Panorama Lite features. Besides offering easy and intuitive usage, this version has better line-chart and graph support and includes tables and subsets similar to the original Panorama. Panorama Lite also features a unique map view that can present data at the county, state, and country levels in a color-coded form so users can visualize geographic patterns relatively easily.

Over the past two years, we've taken this project on the road, collaborating with two organizations to create interactive publications from their traditional, static ones: a publisher (ProQuest) and a government agency (the Centers for Disease Control and Prevention's National Center for Health Statistics, or CDC/NCHS). We created two interactive papers for ProQuest from one of their open-source journals (*Sustainability: Science, Practice and Policy*) that the company launched for public use.

For CDC/NCHS, we converted two issues of their key document — the 2011 and 2012 *Health US In Brief* reports — to interactive form and hosted them on our Web site. *In Brief* contains summary information on the health of the American people, including mortality and life expectancy, morbidity and risk factors such as cigarette smoking and overweight and obesity, access to and use of healthcare, health insurance coverage, supply of healthcare resources, and health expenditures. We also converted another important CDC/NCHS document, the Data Brief No. 115 March 2013, *Death in the United States, 2011*. We expanded Panorama's functionality to support drill-down pie charts for all these interactive papers, and we added a feature to display standard deviation ranges in charts for *In Brief* using HTML5/javascript.

Disaster Information Management: Lost Person Finder

NLM's increasing interest in recent years in mitigating the effects of wide-area disasters has led us to develop several information resources and tools. In our Lost Person Finder (LPF) project, we address the problem of how to reunite families separated by mass casualty events. LPF systems combine image-capture, database, location, and Web technologies, and address both hospital-based and community-wide disaster scenarios.

Web Site and Services

The heart of our system is People Locator® (PL), the main LPF Web site and its MySQL database. We extensively customized the open-source Sahana disaster-management system to create a unified site to hold data from multiple disasters. Missing or found people can be reported to PL by hospital counselors, relief workers, and the public. PL users can report or search for the missing via computer or mobile apps. We made various codebase enhancements, including beginning the process of converting the site to a responsive design framework. During FY2014, we enhanced imaging algorithms and software to reunite families in the wake of disasters, and we added visual search capabilities through use of FaceMatch services.

Deployments

People Locator has been deployed in disasters since the Haiti Earthquake in 2010, as well as in demonstrations and large-scale multi-institutional drills with local Bethesda hospitals. In 2014, PL was deployed for three events: Typhoon Neoguri (July), the Jammu-Kashmir Floods (September), and Typhoon Hagupit (December). In the case of the Jammu-Kashmir Floods in Pakistan and India, more than 10,000 missing-person records were posted to PL (many through Google's Person Finder, which is interoperable with PL). Of these records, about 2,000 had photos of missing people. Additionally, NLM's Specialized Information Services (SIS) used People Locator to demonstrate several disaster-related tools produced by the NLM in the disaster informatics program at the NLM Biomedical Informatics Course held in Young Harris, Georgia (September). Social-networking-project accounts were reviewed and updated for more effective communications. A new social media process and strategy was added to include preparedness messaging for proactive notifications before and during potential disaster events.

Mobile Apps

For hospital-based reporting, the triage process begins with TriagePic®, a Windows application that hospital staff can use to quickly photograph arriving victims. These pictures, along with general health and triage status and minimal descriptive metadata (e.g., name, age range, gender) are packaged and sent by Web services to PL.

For a regional Maryland disaster drill conducted in FY2014 and covering five counties, Suburban Hospital used TriagePic to capture photos and metadata of incoming patients that were searched later for family and friends. We also demonstrated to drill participants the visual search method outlined in the Face-Matching Research section. For community-wide reporting, the ReUnite® app is used and was updated to be compliant with a major mobile operating system release (iOS 8.0+).

Face-Matching Research

Our goal here is to enable users to find missing-person records through automatic face recognition, a significant extension of our current method of searching by name or other text metadata. Our work faces special challenges: unlike many other systems, our face matching has to rely on a single photo of a person to identify her or his face in other images, so we can't exploit traditional face-recognition models that require large sets of photos to train the system.

In 2014, we made considerable progress in face-descriptor matching and retrieval. We researched and developed a new Rotation and Scale Invariant Line-based Color-aware (RSILC) descriptor and confirmed (on the available annotated face datasets) that it is more accurate in matching line-rich objects (e.g., faces) than any individual legacy descriptor used by FaceMatch (FM), but not more accurate than the ensemble of the legacy descriptors. RSILC is currently slower than all legacy FM descriptors or their ensembles, so it needs to be considerably sped up before it is used in the production system. The overall face-descriptor retrieval, however, has been considerably sped up (10 to 20 times faster than in 2013) by using an approach that simultaneously sorts and saves face-descriptor, gender, and age information (at the Web-services level) and the Fast Library for Approximate Nearest Neighbor (FLANN) clustering/indexing (at the FM library level).

In FY2014, our FaceMatch's visual query system performed for our CalTech Faces benchmark dataset at a 98 percent top-1 hit rate, which means that in 98 percent of visual queries, we expect the correct record to appear as the first listing in the results set, with subsecond query turnaround on typical People Locator datasets. The higher the top-1 hit rate, the more reliable a retrieval system is. We increased its top-1 hit rate from 95 percent in FY2013. Our R&D about automatic face-skin-tone clustering has not shown conclusive results yet, and more experiments are needed. PL is currently using both pre-production and production versions of FaceMatch Web services. Current FM services can be used for face and whole-image indexing and searching.

In an effort to make the face retrieval even more robust and efficient, we currently invest in R&D efforts for the person-in-the-loop approach to both face localization and face matching. We collect the problematic images (with false-positive and negative-face detections and misclassifications), analyze our current algorithms' limitations, and investigate incremental learning approaches to skin mapping, face detection, and face matching. To support research and testing, we continue to annotate face images. For this purpose, we developed cross-platform face-annotation tools such as ImageStats (Web-based) and Image List Browser (desktop-based) to gather objective and accurate baseline data. The tools are used successfully by our face-data annotators, such as summer student interns and software testers.

Video Production, Retrieval, and Reuse Project

This development area encompasses projects that contribute to the NLM Long Range Plan goal of promoting health literacy and increasing biomedical understanding. The NLM Media Assets Project gives the NLM easy access to audio-video resources for improved biomedical communications.

Movement Disorders Video Database/MDmedia

This ongoing LHNCBC Research Support Project contributes to improving access to high-quality biomedical-imaging information. We're now able to include video in the clinical study of patients with movement disorders and research the role of mobile technologies in the management of Parkinson's disease by patients and their caregivers. In FY2014, in collaboration with the National Institute of Neurological Disorders and Stroke (NINDS) and the NIH Movement Disorders Clinic, LHNCBC developed a prototype mobile application (for iPad), based on the Movement Disorders Video Database (MDVD), to aid patients with Parkinson's disease and their caregivers in tracking patient's symptoms and medication side effects. The *Movement Disorders Journal* (*MD Journal*) app provides an easy-to-use mobile platform to track speech, dyskinesia, tremor, mood, slowness of movement, and walking and balance on a daily basis. In addition, an interactive log has been developed complete with reminders within the *MD Journal* app that allows patients to track their dyskinesia status every 30 minutes over a 24-hour period. This 24-hour log could replace the paper-based traditional PD diary that patients currently use. In keeping with our initiative to explore the value of video and voice data in mobile devices for patients suffering from movement disorders, these functionalities have been incorporated into the *MD Journal* app. The long term goal is to provide patients with PD and their caregivers with a reliable medical history to be shared with the patient's physicians through a cloud-based or other data sharing system and to be incorporated into the patient's electronic health records (EHRs).

In collaboration with the NIH Movement Disorders Clinic, the *MD Journal* app will be tested by a small group of patients with PD over a 6-month period to determine usability and usefulness of this app in helping to manage patients' disease. After this testing period, user feedback will be collected from the patients, their caregivers, and NIH Movement Disorders Clinic researcher-clinicians to identify areas of strength and weakness within the MD Journal app. Based on this feedback, design and functionality changes will be made for a second deployment and review.

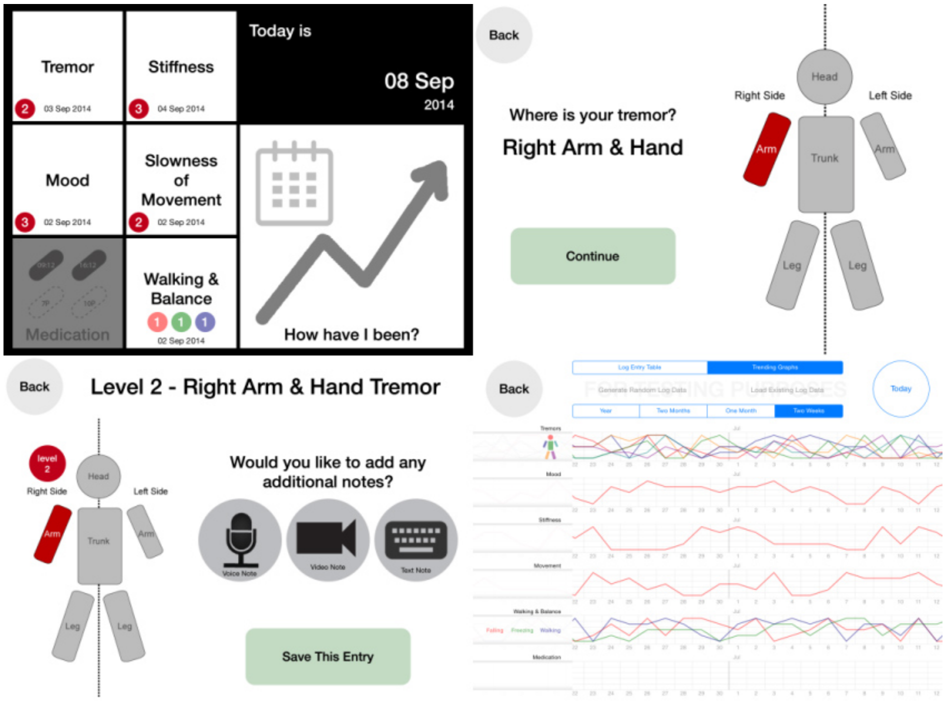


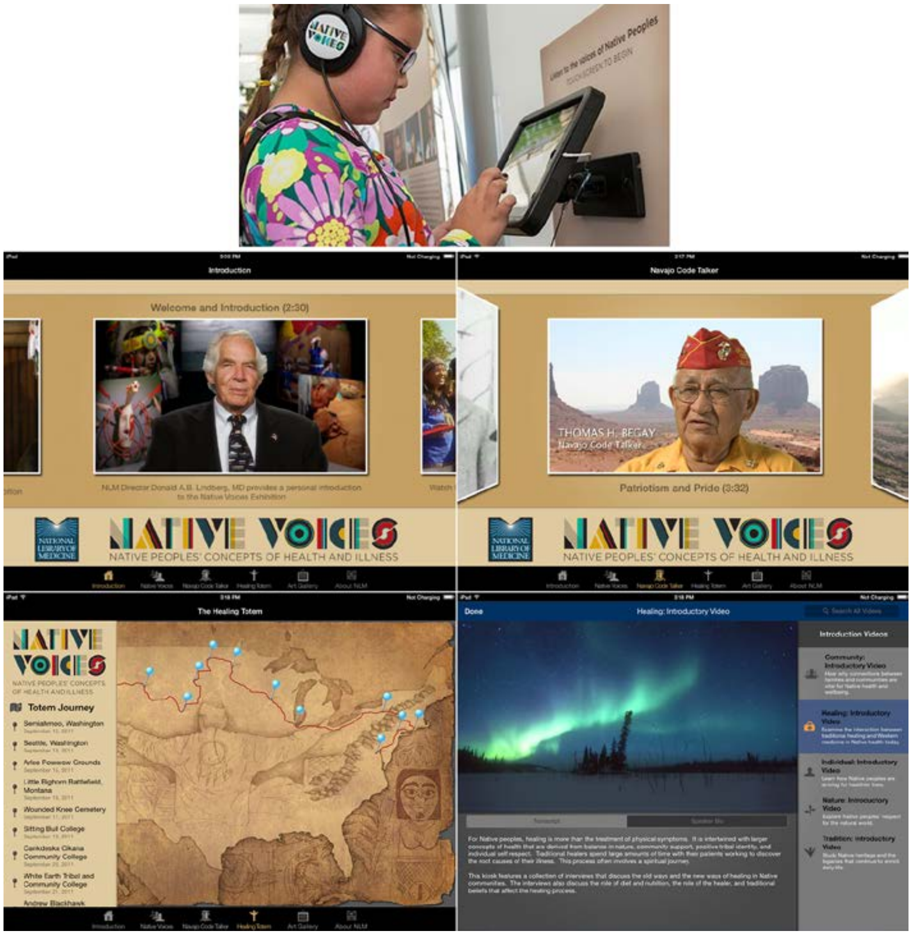
Figure 14. APDB designed a prototype mobile record system which allows patients to track symptoms, medication side effects, moods, and other parameters including personal video recording. NINDS is using this system to test the research hypothesis that patients with Parkinson's Disease can actively manage their care with this app, resulting in improved overall disease management and health outcome. Initial evaluation of mobile app is underway at the NINDS clinic with movement disorders patients and caregivers.

Native Voice Mobile Exhibition Planning, Development, and Deployment

In FY2014, we completed the pilot phase of the Native Voices Mobile Adaptation traveling exhibit, which included successful opening programs and presentations. In collaboration with the NLM Office of the Director, the Office of Health Information Programs Development, the Office of Communications and Public Liaison, Specialized Information Services, and the History of Medicine Division, APDB provided extensive programmatic, technical and logistical planning for all venues, including: the National Congress of the American Indian, Dena'ina Convention Center, Anchorage, AK; The Native American Heritage Center, Anchorage, AK; Southcentral Foundation Native Primary Care Center, Anchorage, AK, Bentuh Nuutah Valley Native Primary Care Center, Wasilla, AK; University of Alaska Anchorage, WWAMI Program, AK; The Queen's Medical Center, Honolulu, HI; John A. Burns School of Medicine (JABSOM), Honolulu, HI; The Hamilton Library at the University of Hawai'i at Manoa; Artesian Art

Gallery, Chickasaw Nation, Sulphur, OK: Texas Medical Center Library, Houston, TX: University of Washington, Seattle.

The traveling exhibition pilot survey instrument was developed for implementation on an iPad Mini, which has been traveling with the mobile exhibition to all venues. Preliminary data has reported outstanding results, with high marks for content, interface, video materials, and overall design. In partnership with the NLM's Regional Medical Libraries, NLM will implement a second phase deployment of the Native Voices traveling exhibition in FY2015.


NATIVE VOICES
Listen to the voices of Native Peoples
TOUCH SCREEN TO BEGIN
Introduction
Welcome and Introduction (2:30)
NLM Director Donald A.B. Lindberg, MD provides a personal introduction to the Native Voices Exhibition
Navajo Code Talker
THOMAS H. BEGAY
Navajo Code Talker
Patriotism and Pride (3:32)
NATIONAL LIBRARY OF MEDICINE
NATIVE VOICES
NATIVE PEOPLES' CONCEPTS OF HEALTH AND ILLNESS
Introduction
Native Voices
Navajo Code Talker
Healing Totem
Art Gallery
About NLM
The Healing Totem
Totem Journey
Semiahmoo, Washington
Seattle, Washington
Arlee Powwow Grounds
Little Bighorn Battlefield, Montana
Wounded Knee Cemetery
Sitting Bull College
Cankdeska Cikana Community College
White Earth Tribal and Community College
Andrew Blackhawk
Done
Healing: Introductory Video
Search All Videos
Introduction Videos
Community: Introductory Video
Healing: Introductory Video
Individual: Introductory Video
Nature: Introductory Video
Tradition: Introductory Video

For Native peoples, healing is more than the treatment of physical symptoms. It is intertwined with larger concepts of health that are derived from balance in nature, community support, positive tribal identity, and individual self respect. Traditional healers spend large amounts of time with their patients working to discover the root causes of their illness. This process often involves a spiritual journey.

This kiosk features a collection of interviews that discuss the old ways and the new ways of healing in Native communities. The interviews also discuss the role of diet and nutrition, the role of the healer, and traditional beliefs that affect the healing process.

Figure 15. *(Top) A girl at the exhibition kiosk uses the Native Voices iPad app. Screenshots from the Native Voices iPad app, showing: (1) NLM Director Donald A.B. Lindberg, MD presents a personal introduction to the Native Voices Exhibition, (2) Navajo Code Talker Thomas H. Begay discusses patriotism and pride, (3) map of the Healing Totem Journey, and (4) Introductory Video about the theme of Native Healing.*

Communication Infrastructure Research and Tools

We perform and support research to develop and advance infrastructure capabilities such as high-speed networks, nomadic computing, network management, wireless access, security, and privacy.

Videoconferencing and Collaboration

We continue to investigate, review, and develop collaboration tools; research their application; and use the tools to support ongoing programs at the NLM. In our work with uncompressed high-definition (HD) video over Internet Protocol (IP), we continued to monitor the HD open-source work of video conferencing tool (VIC) developers on H.264 compression.

Our team also continued investigating some new cloud collaboration tools, including proprietary ones that are standards compliant and that emulate a pioneering collaboration model developed by Argonne National Laboratory. Some commercially available cloud technologies lack features required by NLM, are either too complicated or unstable, or have licensing terms that are not cost-effective given the needs of the Library. We continue to believe that there is no current compelling rationale to migrate from the H.323 standard videoconferencing appliances used for NLM program support. The team also initiated a review of mobile videoconferencing apps. We continued to collaborate with the Rochester Institute of Technology (RIT) and the University of Puerto Rico Medical Campus to test open-source software and commercial cloud technologies for compressed HD videoconferencing based on the H.264 video standard. A manuscript about our cloud findings was accepted for publication, and we are summarizing the results of our mobile app study in a paper we plan to publish in 2015.

We decided to work exclusively with UltraGrid technology for uncompressed high-definition video conferencing because:

- The technology has been enhanced since we started our collaboration with UltraGrid's developers at Masaryk University in Brno, Czech Republic, last year,
- Other uncompressed technologies are receiving less institutional support, and
- The research team at Masaryk is very active and shares our research goals and interests.

In 2014, the team initiated the testing of UltraGrid's extended 4K compressed and uncompressed capabilities, which were added by Masaryk.

A clinical trial of uncompressed video was completed at the Medical University of South Carolina (MUSC), and the equipment there was returned. The trial's purpose is to study the use of uncompressed video as a diagnostic tool. Investigators selected teledermatology as a research focus because previous research showed that it is particularly difficult to use standard-definition video to perform remote dermatological exams. The data-collection phase has ended, and we are now reviewing the data we received and following up on some missing data. Statistical analyses will start soon.

We continued to work with NLM's Specialized Information Services (SIS) on a distance-education outreach program for minority high school students and with the NIH Library to offer distance-education training on the National Center for Biotechnology Information (NCBI) database and other bioinformatics. In FY2014, we conducted bioinformatics programs with the Charles R. Drew University of Medicine and Science, the University of Maryland, the University of North Carolina at Chapel Hill, the University of Tennessee at Memphis, the University of Puerto Rico Medical Campus, and Virginia Commonwealth University. In addition, we began working with the NIH Library to offer distance-learning courses at NIH Institutes located outside Bethesda, especially NIDA in Baltimore and NIEHS in Research Triangle Park, N.C. A high school in Puerto Rico was added to the minority student outreach program with SIS. A manuscript detailing reasons for the distance-education program's success was accepted for publication. Video archives of these and certain other programs are available on the Collab Web server. In 2014, all programs were converted from a Real Media format to one that can be accessed without plug-ins by any HTML 5 — or later — browser. Finally, the team supported a two-day workshop linking, by videoconference, U.S. medical librarians brought to NLM and academicians in Tanzania planning to start biomedical informatics programs. The workshop was sponsored by NLM's International Program.

Collaboratory for High Performance Computing and Communication

LHNCBC established the Collaboratory for High Performance Computing and Communication (Collab) as a resource for researching, testing, and demonstrating the imaging, collaboration, and communications and networking technologies related to NLM's Next Generation Network initiatives and distance-learning research. The Collab serves people both inside and outside NLM.

The facility can be configured for such technologies as 3D interactive imaging (with stereoscopic projection), haptics for surgical planning and distance education, and interactive imaging and communications protocols applicable to telemedicine and distance education involving interactive video and application-sharing tools. These protocols enable staff to collaborate with others at a distance and, at the same time, demonstrate much of the internal and external work being done as part of the NLM Visible Human and advanced networking initiatives. The collaboration technologies include tools built around the H.323 and MPEG video-compression standards for transmitting video over IP and open-source technologies such as Conference-XP, iHDTV, UltraGrid, and the Access Grid, as well as cloud technologies and mobile apps.

In addition to supporting research, the Collab is a vehicle for offering the SIS distance-learning program, biotechnology training, virtual site visits, and other virtual meetings. We archived some programs on the Collab Web server.

Computing Resources Projects

We conduct numerous projects to build, administer, support, and maintain an integrated and secure infrastructure to facilitate LHNCBC's R&D activities. The integrated secure infrastructure contains network, security, data storage, and facility management, as well as system administration support for a large number of individual workstations and shared servers.

In FY2014, to comply with Key Federal IT modernization initiatives, we successfully deployed IPv6 to desktops and public services, including domain name system (DNS) and Web services. We achieved a 90 percent Federal Desktop Core Configuration (FDCC) compliance rate in FDCC and U.S. Government Configuration Baseline (USGCB) (10 percent more than the NIH average). Following the guideline of Key Federal IT modernization initiatives in cloud computing, we successfully deployed remote-access thin clients, computers that don't require typical desktop configuration because they connect to a main server and don't include a hard drive. The thin clients were 50 percent more cost-effective than regular desktops, they enhanced security measures, and they reduced maintenance efforts.

We reduced LHNCBC's software maintenance cost by 13 percent by consolidating RedHat Licenses and renegotiating an extended network-storage maintenance contract. To ensure Continuity of Operations (COOP) per the National Security Presidential Directive, we established a COOP Subject Matter Expert (SME) list with 24/7 monitor, process, and -recover procedure to ensure availability of LHNCBC network, security, centralized data storage, remote access, and critical public services.

Training and Education at LHNCBC

LHNCBC is a major contributor to the training of future scientists and provides training for postdoctoral fellows and other people beginning their biomedical research careers. Our Medical Informatics Training Program (MITP), ranging from a few months to two years or more, is available for visiting scientists, postdoctoral fellows, and graduate and medical students. Each participant spends between a few months and several years working on LHNCBC research projects under the guidance of 16 LHNCBC research staff mentors. A list of MITP frequently asked questions (FAQs) on our Web site include information about the lecture series that is part of this program. Participants also make presentations, write manuscripts, attend and present at professional conferences, and may publish in professional journals.

During FY2014, 40 trainees (including 19 postdoctoral fellows) from 18 states and 7 foreign countries received training and conducted research at LHNCBC in a wide range of disciplines, including:

- 3D image processing,
- Big data to knowledge (BD2K) based on large biomedical datasets,
- Biomedical ontology and terminology,
- Clinical health information question answering (CHIQA) systems,
- Collaboration tools,
- Content-based information retrieval,
- De-identification of medical records,
- Evidence-based medicine systems,
- Information retrieval,
- Literature-based discovery,
- mHealth, image, text, and document processing,
- Natural language processing,
- Personal health records,
- Pill identification,
- Semantics Web research, and
- Disaster management.

We emphasize diversity by participating in programs for minority students, including the summer internship programs of the Hispanic Association of Colleges and Universities and the National Association for Equal Opportunity in Higher Education.

The MITP sponsors the Clinical Informatics Postdoctoral Fellowship Program, funded by LHNCBC, to attract young physicians to NIH to pursue research in informatics. This program is run jointly with the Clinical Center to bring postdoctoral fellows to labs throughout NIH. We continue to offer an NIH Clinical Elective Program in Medical Informatics for third- and fourth-year medical and dental students, which offers students the opportunity for independent research under the mentorship of expert NIH researchers. We also host a two-month NLM Rotation Program that gives trainees from NLM-funded medical informatics programs an opportunity to learn about NLM programs and current LHNCBC research. The rotation includes a series of lectures showcasing research conducted at NLM and provides an opportunity for trainees to work closely with established scientists and fellows from other NLM-funded programs. We also provide an informatics lecture series and submit project proposals for the NLM Library Associates program.