Picture archiving and communication systems (PACSs) have been implemented in the last decade to streamline radiologic image management, and the conversion to “filmless” radiology departments is accelerating. Many large PACSs also support the archival of textual radiologic reports in raw file format for display together with images in the PACS review workstations. The Digital Imaging and Communications in Medicine (DICOM) Standards Committee has recently approved a new class of structured objects dealing with the generation, distribution, and management of textual reports. The structured reporting (SR) objects bridge the traditional separation between imaging and information systems. The DICOM SR objects offer a higher level of integration of the medical imaging enterprise, providing practitioners with an effective tool to cover many aspects of the medical imaging services process, from admission to discharge. With the approval of DICOM SR standards, it is expected that PACSs will gradually become the primary source of radiologic reports as well.

With rapid advances in basic scientific and imaging techniques, clinicians and scientists now collect complex data in ever-increasing amounts, fostering increased specialization, with resultant challenges to integrate data between and across levels of interaction, control, and function. The sheer quantity and complexity of the data are such that the field of neuroscience, as well as that of imaging research, would...
benefit considerably from an information management system approach for its experimental and clinical data. Both fields should enhance their wealth of ever-increasing empirical data, accumulated from their many disciplines and experimental approaches, by developing appropriate databases and associated tools as well as a greater capability for both theory development and simulation models.

PACSs lack sophisticated indexing and processing capabilities for managing or mining the data derived from image and data analysis for a broader range of clinical services and scientific research. We have been steadily working on various aspects of image database system research for the past decade. In this paper, we describe a neuroimaging data warehouse framework that models and integrates multiple neuroimaging modalities, including MRI (magnetic resonance imaging), x-ray CT (computed tomography), PET (positron emission tomography), MRS (magnetic resonance spectroscopy), and associated clinical and research data to support neurologic diagnosis and neuroimaging research.

The design of the framework is based on object-oriented analysis and design (OOAD) methodology and recognized industrial standards, such as HL7 (Health Level 7) and DICOM, in retrieving data from underlying clinical systems. The OOAD method includes the Unified Modeling Language (UML), use case-driven design, and distributed object computing technology, i.e., CORBA (Common Object Request Broker Architecture). The image data warehouse system provides more effective utilization and analysis of multimodality images and records in supporting clinical research and care than what is currently supported by PACSs, hospital information systems, or other medical databases in silos.

**Background**

**Picture Archiving Communication Systems**

To meet the challenges of acquiring, archiving, communicating, and displaying medical images, PACSs were developed during the last decade as image management systems. First-generation PACSs were designed to support direct radiologic softcopy readings but were not intended to integrate other related textual information sources of radiologic practice. Several hundreds of the first-generation systems are currently deployed in hospitals worldwide.

The subsequent development of second-generation PACSs, also known as hospital integrated PACSs (HI-PACSs) or enterprise PACSs, included the integration of PACSs with heterogeneous information systems like hospital information system (HIS) and radiologic information system (RIS), the migration to an open client/server system architecture, and the implementation of recognized industrial standards in image and data format, i.e., DICOM and HL7. This integration trend continues as PACSs take their place in the enterprise effort to provide more efficient and cost-effective access to all medical data to support patient care. Hospital integrated PACSs are more expensive to deploy, usually costing a few million to tens of millions dollars.

PACSs are being widely implemented in radiology departments today to support daily clinical diagnosis. The systems are adequate to facilitate diagnostic softcopy reading, but they fall short of supporting other kinds of image and textual data services outside the radiology department. The recent approval of DICOM SR will enable users to archive and query diagnostic reports directly from PACSs and also provide means to better integrate other medical reports. This will potentially provide a more efficient and consolidated clinical data store to feed into image data warehouse servers.

**Extending Imaging Services Outside Radiology Reading**

Clinicians and researchers outside radiology departments want to have as easy access to the medical images stored in PACSs as do radiologists themselves. Currently deployed PACS display workstations are expensive, have limited image processing capabilities, and have no utilities for storing results of analysis besides textual diagnostic reports. These display workstations are also heavily used by radiologists and are difficult to access by researchers and clinicians who are not conducting routine diagnostic readouts. Although the operation of PACSs has led to the accumulation of vast stores of multimedia medical data files, including raw images, patient data, and diagnostic reports, accessing and analyzing these images outside the scope of daily clinical operation is a long and tedious process.

What is the current practice of analyzing images for research and diagnosis? Clinicians and biomedical researchers employ powerful computer workstations to retrieve digital images from a PACS and other image sources, store images on local disks, visualize multiple image sets simultaneously, and apply image-processing routines such as segmentation, quantitation, and registration. All the patient images
and text data derived from such analysis are stored on the local disk.

The drawbacks of using such computer workstations are many. The data are inefficiently managed via flat files on hard disk, and there is no redundancy to ensure data integrity. The analysis work must be done on a single machine; collaborative research between multiple individuals is inhibited. The workstations are expensive and not easily accessible.

Such obstacles hinder the ability of clinicians outside a radiology department to collect and analyze medical images when preparing diagnostic workups of patients for epilepsy surgery and determining decision thresholds. For example, in our initial selected application area of data warehouse for seizure disorders, the Northern California Comprehensive Epilepsy Center of the University of California–San Francisco (UCSF) and the Epilepsy Center of University of Alabama utilize several neuroimaging modalities as well as other textual and video signal data in the diagnosis of patients with intractable temporal lobe epilepsy. Neurologists, neuroradiologists, neuropsychologists, and neuroscientists spend a great deal of time, often weeks, gathering data from disparate information systems and paper records before sitting down to do a diagnostic workup or data analysis.

**Design Objectives**

This informatics project applies industrial standards and methods to development of an image data warehouse framework for multimedia management, data analysis, research, and access services. The framework is a complete description of the problem that abstracts the many aspects of epilepsy research and care to a high level of visual understanding. It can be modified and expanded to provide new services or support new application domains. The design patterns and modular architecture used in the framework will be useful in solving problems in different medical domains. The entire design process is use case driven.

**Image Data Warehousing Approach**

Up to now, the primary purpose of most database systems was to meet the needs of operational systems, which are typically transactional in nature. Hospitals today generate and collect vast amounts of data in the ongoing process of providing care. Classic examples of operational systems in hospitals include patient registration and order entry (HIS), image reports and scheduling (RIS), and image display and reading (PACS). Operational systems by nature are concerned primarily with the handling of a single transaction and are optimized and normalized for transactional updates. Normalization is a term for a process the data modeler goes through to avoid the traps and pitfalls easily encountered when the network of entities, attributes, and relationships that make up a relational data model that is commonly used in today’s database systems. Most of the time, the needs of an operational system do not change much.

On the other hand, a data warehouse application is different, as a typical transaction deals with large amounts of data, which are aggregate in nature. The operational entities are often denormalized in the data warehouse, a process in which the structure of the tables from the operational environment is changed to allow faster processing in the data warehouse. The data stored in the data warehouse undergo little if any update activity. Their main purpose is to be read by users while they make decisions or verify hypotheses.

Under today’s competitive managed care environment, health care decision makers must be able to interpret trends, identify factors, and utilize information based on clear, timely data presented in a meaningful format. Biomedical researchers would use the data warehouse to verify scientific hypotheses and deal with “What if?” questions in exploring various aspects of disease management and treatment.

**Figure 1** shows the process flow for developing medical image data warehouses. The data warehousing process consists of five elements—prepare the data, acquire and load the data into a data staging area, transform and extract the data, archive the data into the data warehouse servers, and analyze and mine the data. The major tasks in each process element are described in the figure.

Several features distinguish medical image warehousing systems from the data warehousing systems used in other business organizations or industries:

- A medical imaging data warehouse would be built more conveniently in a PACS or digital imaging department environment in which massive volumes of images and associated reports are acquired and archived centrally. The availability of message exchange standards and software interfaces such as DICOM and HL7 also help in reducing the complexity of data preparation and acquisition.
A medical imaging data warehouse involves sophisticated effort of image processing, registration, extraction, and quantitation of image features. It also requires the modeling of multimedia data in various forms—including free text, structured reports, planar and volumetric images, zoomable images, signals, spectroscopic data, graphics, video clips, and scanned paper records—rather than simply textual data, as in most business data warehousing systems. The ability to search all forms of information, qualitatively and quantitatively, not just structured text and numbers in records, makes the image data warehouse a true enterprise information portal.

An image data warehouse emphasizes the preparation and acquisition of data with predefined or prospective protocols, rather than retrospective analysis or mining of data. The emphasis is due to the predominant paradigm adopted in biomedical research today.

An image data warehouse provides analytic and statistical tools to support a verification-based approach in which the user hypothesizes about specific data relationships and then uses the tools to verify or refute the hypotheses. This follows the hypothesis-driven paradigm adopted in most biomedical research endeavors. Data mining, in contrast, uses what are called discovery-based approaches, in which pattern matching and other algorithms are employed to determine key relationships in the data. The ability to automatically discover important information hidden in the data and then present it in the appropriate format is a critical complementary technology to verification-based approaches. Since the goals are exploratory in nature, a data mining system would normally require a much larger data size than a data analysis system. In addition, discovery-based tools and techniques are relatively immature, compared with verification-based tools, and are often a research topic of their own.

Iterative Analysis and Design Process

The differences between a data warehouse and an operational information system occur at the very fabric by which these technologies are woven. System design life cycle is the common methodology used to build any operational system.11 For most operational systems, we can gather the correct requirements with proper analysis in the beginning. Once we have this analysis, we can begin to construct the development plan. We then model the operational system. A high likelihood exists that the model we develop will become the finished product, unless a major mistake is discovered during the development. We continue with this process until the system is complete. Typical enhancements are adding simple new features or changing reports. We would not rethink the core functionality of the system. Because the requirements are clear, the process does not need to be iterative to be successful.

With a data warehouse project, it is difficult, if not impossible, to do anything other than iterative analysis. Because the end users often do not yet know how they will analyze the data, it is impossible to start by gathering all requirements. The goal of a data warehouse is to be flexible enough to deal with the changing needs of the business so that as we ask “what if?”
questions, we can mine deeper and deeper into the warehouse. Each question probes and interrogates the underlying information in new ways. If this is done correctly, we will find the answers we seek.

Henceforth, in our image data warehousing effort, an OOAD process is followed to create a logical and iterative model of the software solution.12 This process transforms real-world concepts—such as medical images, clinical records, laboratory test results, phenotype and genotype information, physicians, neuroscientists, and patients—and transforms them into a logical model on paper. Use cases are created to summarize operational scenarios of clinicians using the system to complete certain tasks, such as image registration.13 After the use cases are prioritized, a spiral development process is followed to incrementally build computer services to support each use case. The spiral development process consists of multiple development cycles, each of which has its own analysis, design, and construct phases.

The Unified Modeling Language (UML) is used to describe the object-oriented concepts of the system throughout the software analysis and design process. The UML is a formal language for describing object-oriented concepts and relationships. The language is visual and communicates everything from the requirement analysis to methods and objects to be implemented in code. This means that throughout the analysis and development process, the system documentation is generated in parallel with the artifacts of the process. Figure 2 illustrates the important UML building blocks and how they flow into the technical documentation.

The advantages of using a formal language such as UML are support for the iterative development process and creation of encapsulated software modules that can be re-used in the future. The UML has the advantage of existing independently of any programming language and, arguably, of any implementation methodology. Therefore, the logical model described by UML can be re-used as technologies advance.

Another advantage of UML is that it creates a communication standard and a consistent method of create incremental documentation.14 For example, the health care information standards, such as HL7 version 3 and CORBA-MED, adopt the UML as a formal language for modeling.

A use case is a narrative document that describes how an external actor uses the system to complete a specific task. The domain models are constructed on the basis of use cases to illustrate how the system would function on a logical level in order to complete all the tasks specified by the use cases. A use case is written in plain English so that the user (e.g., a clinician or an imaging specialist) and the designer can agree on how the system will solve the problem at hand. The use case not only serves as the basis for design but also is used for evaluating how well the system fulfills customer expectations.

In addition, one can draw association relationships between use cases and use cases. For example, a generalized relationship between use cases shows that a use case can share the behavior defined in one or more use cases. Figure 3 provides an example of a leaf use case, which is an end node of a use case diagram.

Once all the use cases have been determined, a use case diagram is generated to show the relationship between the external actor and the use cases. In Figure 4, for example, the clinician user interacts with one of four use cases— visualize image data set, register two image data sets, segment image features, and visualize lab test results. The two use cases visualize image data set and visualize lab test results access the use case validate user.

The design phase develops a logical solution based on the object-oriented paradigm. The main task of this phase is to create interaction diagrams that illustrate how objects will communicate in order to fulfill the requirements. The creation of interaction diagrams requires the application of principles for assigning responsibilities and the use of design patterns.
The single most important step in OOAD is the judicious assignment of responsibilities to software components, and is usually accomplished with design patterns. A design pattern is a named description of a problem, a solution, when to apply the solution, and how to apply the solution in new contexts. Skilful use of design patterns can improve the understandability, reusability, and flexibility of a software system.

With the completion of interaction diagrams, the last step in the design phase is translating the interaction diagrams and conceptual model into software classes and methods. Design class diagrams are created to illustrate the specifications for software classes and interfaces in an application. In contrast to a conceptual model, a design class diagram illustrates software entities rather than real-world concepts.

The design class diagrams are mapped into an object-oriented programming language by the writing of source code for class definitions and method definitions. A good development effort is characterized by a significant amount of analysis and design modeling before coding begins. Rushing to code creates systems that are harder to understand, extend, and maintain and does not support a successfully repeatable process. However, unexpected insights may be gained during the construct phase and will be inte-
grated in the next development cycle of the spiral development process.

When an expert has fine-tuned a solution to solve a particular design problem, he or she will use it again and again in the future. Thus, object-oriented systems contain many recurring patterns of classes and communicating objects. These design patterns solve specific problems and make object-oriented designs more flexible, elegant, and ultimately reusable. They allow designers to create successful systems based on prior experience. The smart designer knows that it is not always smart to solve a problem from first principles when the problem has already been solved with a simple and proven design pattern.

Design patterns make it easier for a designer to reuse successful designs and architectures. By documenting such successful solutions, the patterns can be applied in new contexts by designers of new systems. Design patterns not only help the designer make the right choices in creating reusable, efficient systems; they also help a designer document and maintain an existing system by providing a vocabulary for describing the system in terms of its underlying patterns. Especially in large software systems, the number of objects, relationships, and interactions can be overwhelming, and patterns simplify the system description, because the seasoned designer will comprehend the functionality based on the design pattern.

**System Description**

**Image Data Warehouse Architecture**

The multimodality image data warehouse framework is based on a multitiered architecture that separates the graphical user interface (GUI) applications, data warehouse business service components, data staging area, and backend data source systems into distinct layers. Figure 5 show the testbed implementation of the UCSF neuroimaging data warehouse system.

**Source Systems**

Source systems include various clinical operational systems and research databases—such as a PACS, RIS, HIS, neurology database, neurosurgery clinics database, pathology laboratory database, and psychology research database—from various research laboratories and clinical sections of UCSF hospitals in
the San Francisco Bay Area via secured broadband ATM (asynchronous transfer mode) campus intranet.

The data in our image data warehouse system can be broadly categorized into neuroimaging modalities, digital microscopic images, and findings from neuropathology, diagnostic reports, psychometric tests, patient demographics, laboratory tests, surgical findings, and postoperative outcomes 6 months, 1 year, and 2 years after surgery.

Patients undergo a series of examinations under well-defined protocols at the Northern California Comprehensive Epilepsy Center and the University of Alabama Epilepsy Center. Figure 6 outlines the general diagnostic protocol for presurgical epilepsy diagnosis. Patients entering this protocol must satisfy one of two conditions: 1) they are medically refractory; that is, their seizures are not controllable by anti-epileptic drugs, or 2) they experience intolerable side effects from anti-epileptic drugs.

The goal is to determine whether a patient is a good candidate for surgery. In good candidates for surgery, the epileptogenic focus has been localized from the results of various clinical, electroencephalographic, and neuroimaging examinations. The successful finding of concordance among the noninvasive imaging modalities—e.g., MRI, PET, MR spectroscopy imaging, MR perfusion imaging, MEG (magnetoencephalography)—or a combination of two or more modalities, precludes the performance of invasive intracranial electroencephalography (EEG) monitoring from surgically implanted electrodes.16

Each brain image data set contains a DICOM header, which itself contains textual data regarding the patient’s name, imaging modality, scanning parameters, date, attending physician, and such. The extraction and composition of textual data from the diagnostic reports and the DICOM header fields can be automatic, whereas the segmentation and extraction of medical images often are done interactively. Since imaging vendors may choose to select certain proprietary sections of the DICOM header (e.g., odd group number instead of standard even group number) to store information unique to their applications, prior arrangements must be made with vendors to ensure the inclusion of such proprietary information in the image data warehouse. For our applications, the DICOM interface is developed on top of the Mallinckrodt Institute of Radiology’s Central Test Note public domain software.

The clinical and laboratory tests include complete blood count, electrolyte panel, liver function enzymes, thyroid function, and drug medication tests. The test results are retrieved from the online HIS via an HL7 gateway. Software modules extract the test results from the HIS reports and insert them into the epilepsy data warehouse. The HL7 interface is based on the commercial interface engine, Symphonia3 (Orion Inc., Auckland, New Zealand).

The neurosurgery and neurology databases reside on separate PCs running Microsoft Access in the campus-wide network. The data are first dumped into Microsoft Excel files and then automatically loaded into the epilepsy data warehouse. The psychometric test results are manually entered into the database via the Web from the paper records, whereas pathology findings and images are obtained by FTP transfer.

Data Staging Area Servers

Data staging area is the initial storage and cleaning system for data that are moving toward the data warehouse server. They consist of a set of computational processes that clean, transform, combine, deduplicate, prepare, and export source data for model and use in the image data warehouse. It would be nice if the data staging area were a single centralized facility on one piece of hardware. In practice, since data come from heterogeneous source systems on different platforms and because of the need to use a
variety of software packages for various image processing tasks, the data staging area of the data warehouse is spread over a number of computers (a Dell 2300 dual processor NT server, SUN Solaris E2 server, and SGI Onyx IRIX server).

CORBA is used as the underlying distributed object technology for integrating data of source systems via component wrapping of DICOM and HL7 interfaces and for enabling the communication among software modules residing in the data staging servers. CORBA is a popular and standardized component-based architecture developed by the Object Management Group. Its architecture consists of interoperating software components that provide access to services such as laboratory data retrieval or DICOM image query.

In our data warehousing project, we use a commercial Visibroker product (Borland Software Corporation, Scotts Valley, California). The image processing components of the data staging servers provide and select image processing algorithms for image filtering, coregistration, enhancement, feature extraction, quantitation, and structural/functional mapping. A number of clinical data retrieval components provide access to laboratory test results and patient records from the HIS and other specialty database systems.

The image processing components in the data staging area are implemented on a high-end Silicon Graphics Onyx series computer (Mountain View, California) and an Apple PowerMac G4 (Cupertino, California) running the BrainImage software from Stanford University. The image registration software, Automated Image Registration (AIR) from UCLA, runs in a Solaris environment on a Sun Ultra10 workstation. The PACS DICOM gateway and the HIS HL7 gateway, as well as their CORBA wrappers, reside on a SUN Enterprise 2 computer and a Dell 2300 PowerEdge NT computer, respectively. The network consists of 10 and 100 BaseT Ethernets. The data staging area also provides temporary storage in both a relational database (Oracle 8.0.5) and a system flat file disk space of 30 GB to clean, transform, combine, and prepare source data and images files.

For our current image data warehouse project, dedicated resources (research assistants) have been employed to perform routine and defined data loading and transformation operations. They are willing to participate since they would get values out of the integrated data warehouse that they could not do so alone.

One key defining restriction on the data staging area is that it does not provide query and presentation services. As soon as a system provides query and presentation services, it must be categorized as a data warehouse server, which is described next.

Data Warehouse Servers

The target physical machine is a Dell 6500 dual processor NT server on which the data warehouse data are organized and stored for direct querying and analysis by end users, researchers, clinicians, and applications.

The data warehouse layer includes the several application servers and business service components. The application servers include the Web application server, image processing server, and image database server. The epilepsy data warehouse stores and represents data of quantitated image features, various instance-UIDs (e.g., StudyinstanceUID, SeriesinstanceUID, SOPinstanceUID), keywords, thumbnail pictures, and annotated images into an object-relational database (Oracle 8i) and raw DICOM or processed images of multimodality in a RAID 5 system.

The images include MR images of various protocols and MRS images. PET and magnetoencephalography (MEG) images are also included but are not archived in the UCSF PACS. The business service components manage, integrate, and distribute the data to a group of clinicians, neuroimaging specialists, and neuroscientists working in concert on complicated research and clinical projects. The functions of each service component are briefly as follows:

- The knowledge base component controls the high-level interaction between the user and the servers, encodes medical rules for decision aids, and provides interactive workup models of different epilepsy categories.
- The user profiling and authentication component manages the customizable user interface for individual clinicians, depending on their personal preferences. After requests for information have been processed by other middleware components, the processed data are sent to the user profile component for formatting before they are sent to the presentation layer. The service component also provides user authentication and audit trail capabilities.
The information retrieval component supports ad hoc user query and keeps information on image classes, search algorithms, image features, keywords, pictures of interest, and their association to support content-based retrieval of medical images.

The image visualization component supports the viewing, mapping, and manipulation of three-dimensional medical images on the Web.

The data analysis component contains algorithms for searching the data warehouse and linking external commercial statistical packages, such as SPSS, to analyze image data and associated clinical and research findings. For example, the user can invoke the service component to calculate and analyze clinical thresholds for structural volume and glucose metabolism that indicate lateralization of the epileptogenic focus.

The image data warehouse is implemented on a dual-Pentium processor Dell 6500 PowerEdge server running the Windows NT 5.0 Server operating system and Oracle 8.1.7 database management system, which is clustered with another dual-CPU Dell 2300 PowerEdge server for backup purposes. A database management system provides centralized storage of all multimedia neurologic data so that clinicians are not tied to particular workstation and so that the data of multiple clinicians working together may be managed, integrated, and distributed. The medical images are stored in both DICOM and raw volumetric binary data formats.

Costs and development efforts have been saved by utilizing existing computing resources more efficiently and by moving the user interface to less expensive personal computers while maintaining accessibility to more powerful workstation servers in the middleware layer. These servers that provide CPU-intensive services such as image processing and data analysis functions. The data warehouse front-end includes both Web browser access and a dedicated multimedia GUI client on either a Unix or PC platform. The current setup is to have the Web browsers access the image data warehouse via intranet or HTTP over SSL (secured socket layer) for users outside UCSF campuses.

Registration and Quantitation of Image Data

A key feature distinguishing an image data warehousing system from a conventional data warehousing system is the ability of the former to manipulate, query, and correlate quantitative image data. Figure 7 shows the processing steps that occur in the data staging area to extract quantitated brain image and textual features from the operational systems' store data, e.g., from the PACS, RIS, and HIS data. The processing steps can be generalized for other imaging modalities.

Feature extraction is based on the a priori approach, in contrast to dynamic and automatic feature extraction during user query, as proposed in many nonmedical image database applications. For image data warehousing systems, coregistration of functional PET and T1-weighted MR images is an essential step toward combining functional information from PET images with anatomic information in MR images.

Several coregistration algorithms have been published in the literature and are used in functional brain imaging studies. Image registration permits the combination of different types of functional information (such as PET and MEG images) and structural information (such as MR images), setting the stage for feature extraction. Our information system interfaces with the public-domain Automated Image Registration package (AIR, version 2.03) as well as several proprietary routines in commercial image analysis packages for registration of neuroimages, including PET, MRI, MRS, and MEG.

The correlated image data sets are encoded into the targeted data warehouse model to serve as definitive indexing in image database queries. For example, registering the functional images of PET with MR images of the same patient allows the intrinsically better spatial resolution of MR images to be used in quantitatively analyzing functional information (metabolic count of glucose consumption) of captured PET scans. In epilepsy surgical evaluation, neurologists are able to pinpoint those parts of the brain with suspicious hypometabolic activity and compare their locations with the geometric location of neuronal cell loss in the MR images. The regions of interest derived from segmenting the MR images are transformed into the PET space for comparison.

The segmentation and extraction of medical images are done interactively using the volumetric image display and analysis (VIDA) environment and BrainImage software. Image features are stratified into different levels of detail: Image data set → anatomic structures → pathologic → microscopic → genetic. We divide the image features into primitive and logical. Primitive features are directly obtained from the medical images, such as volume, shape, and texture of certain organs in MR images or metabolic activities of brain tissue in PET scans. Regions of interest are semi-automatically outlined in the MR images to obtain
anatomic volumes, and coregistration with functional images allows quantitative analysis of functionality in the corresponding anatomic region.

Logical features are abstract representations of images at various levels of detail and deeper domain semantics that indicate, for example, whether the volume of an anatomic structure is normal or whether certain brain tissue is hypometabolic in reference to certain established data. These logical features are synthesized from primitive ones and additional domain knowledge.

All extracted features are entered into appropriate attributes of the objects defined in a data model, to facilitate subsequent information query by content. For seizure disorder applications, we have extracted MRI anatomic volume, PET glucose uptake count, MRS spectra, and MEG dipole polarization for the amygdala and hippocampus. We have also gathered pathologic samples from surgery and digitized them using a SPOT II digital camera. We are thus able to examine brain structure and function from several levels of detail for clinical research and outcome analysis.

The extraction and composition of textual data from the diagnostic reports and DICOM header fields can be automatic. Certain keywords or phrases are automatically extracted from the physician textual reports for indexing purposes. All commercial PACS image files have DICOM headers, which contain patient and imaging examination information. The DICOM header is organized into sequential data element tags, each consisting of a group number and element.
number. For example, the Patient Name value is located in group 0010, element 0010 (Table 1). This value is automatically extracted and entered into the data warehouse column Patient’s Name.

**Status Report**

We are utilizing the neuroimaging data warehouse system to support several clinical care and research areas of seizure disorders. This section briefly illustrates two distinct usages of image data warehousing systems in supporting clinical practice and medical research.

**Clinical Practice Example:**

**Diagnostic Workup of Neuroimaging Cases**

Image diagnostic workup is often a time-consuming and laborious endeavor that involves gathering all relevant medical records and images from various information systems and paper filing systems. The collection and preparation of a diagnostic workup evaluation can sometimes take weeks. The availability of an integrated data warehouse system that captures patient records and images for particular diseases provides a one-stop solution for faster and more effective diagnostic workup.

When a clinician sits down at a client computer to initiate an analysis session, the first step is to retrieve image and textual data from the PACS and other data sources. Figure 8 shows a screen that appears during the retrieval of image and textual information from the data warehouse to the desktop workstation. The GUI was developed in the GainMomentum multimedia development environment (Sybase Inc., Emeryville, California) using proprietary scripting language.

In the upper left of the screen, the patient’s name is entered (the patient name and ID have been changed to maintain privacy). The system returns descriptions of the imaging studies corresponding to the specified patient name, and these items are displayed as rows in the upper scroll widget. Clicking on a row triggers retrieval of the diagnostic report, and clicking on the “Retrieve Image” button triggers the retrieval of the selected image data set. Software modules automatically convert the DICOM-based image file to formats readable by the visualization packages on the desktop workstation.

![Figure 8](http://jamia.oxfordjournals.org/) An example of a desktop multimedia graphical user interface for information retrieval in the medical image data warehouse.

### Table 1

<table>
<thead>
<tr>
<th>Group</th>
<th>Element</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0010</td>
<td>0010</td>
<td>Patient Name</td>
</tr>
<tr>
<td>0010</td>
<td>0020</td>
<td>Patient ID</td>
</tr>
<tr>
<td>0018</td>
<td>0015</td>
<td>Body Part Examined</td>
</tr>
</tbody>
</table>

---

WONG ET AL., Image Data Warehouse Framework
With the imaging studies now residing on the workstation, biomedical information is quantitated, extracted, and combined in various ways to aid in the localization process. First, the non-brain structures, such as scalp and skull, are automatically segmented from the MRI data set, and a transformation matrix is generated to coregister the volumetric PET with the newly processed MRI data set. The user can set several registration parameters, such as sampling, iterations, thresholds, and initialization files.

Exploiting the excellent anatomic details of MR images, the region-of-interest utilities of the neuroimaging data warehouse system are used to interactively segment the hippocampus and amygdala. The resulting segmentation parameters are combined with the transformation matrix and PET sinogram files to yield glucose counts for these mesial temporal structures (Figure 9). Such quantitative data, volumes and glucose counts, cannot be obtained by conventional workups and are important in determining the suitability of patients for surgical treatment.

After the workup, the clinician saves all the analyzed results into the epilepsy data warehouse for future studies. These include not only the final numeric values of volume, glucose counts, electrical spike magnitude and location, and metabolite concentration but also the segmentation parameters and transformation matrix that were generated at the workstation.

The power of organizing the analyzed data in the data warehouse is the ability to do online analytic processing of the collected data. The database can be queried on any number of keys, ranging from patient name to image features and report keywords. Complex queries—such as “Find female patients age over 21 years with right hippocampal atrophy > 10% and corresponding depletion of N-acetyl-aspartate > 15%” can also be executed. Names of matching patients are returned to the workstation, and full image data sets can be retrieved by clicking on the thumbnail images. It is important to note the ability to query the quantitative values of image features using the image data warehouse.

**Medical Research Example:**
**Data Analysis and Decision Threshold on Seizure Foci Lateralization**

Electroencephalography is the most common method of localizing the epileptogenic zone. Nevertheless, paroxysmal abnormalities occur intermittently and may not be easily captured. Inpatient recording of scalp EEG with video is necessary to capture the seizures. Magnetic resonance imaging examinations, and sometimes PET scans, are used to provide additional information about localization of an epileptogenic zone.

The Northern California Epilepsy Center currently performs additional noninvasive neuroimaging examinations—i.e., PET, MRSI, and MEG—related to clinical research, on a selected group of epilepsy patients. The multiple-image data sets from these patients are processed, organized, and modeled for online data interpretation and analysis. The use of combined noninvasive tests to eliminate or reduce the frequency of intracranial EEG recordings will help improve patient care, minimize complications of
invasive tests, reduce costs, and shorten hospital stays. In addition, improved accuracy of epilepsy localization will increase the number of medically refractory cases of epilepsy that are remediable by surgery treatment.

The neuroimaging data warehouse framework has been utilized to analyze brain images from 80 patients. In cases in which no hippocampal abnormalities were detected on MR images, coregistered MR images were used to detect subtle lateralized abnormalities on PET images. Missed abnormalities and false interpretations of independently interpreted high-resolution F-18-fluorodeoxyglucose (18FDG)-PET images were correctly identified with coregistration to MRI, allowing detection of hypometabolism in patients with partial epilepsy of mesial and neocortical temporal origin.

The data analysis module of the image data warehouse is also used to investigate the lateralization concordance between various neuroimaging modalities in epilepsy. Such types of queries are valuable for determining the clinical efficacy of the diagnostic procedure. Figure 10 shows sample data analysis pages. In this application, surgery result is used as the gold standard for reference, and MRI and scalp EEG predictions are compared with each other and
with surgery results in patients who underwent surgical treatment more than a year previously.

The data warehouse allows users to customize the presentation format (personalize the display layout) beforehand. In Figure 10, the user has selected pie charts and bar graphs to illustrate the analyzed results, instead of tables or text formats. Furthermore, as shown in Figure 10, bottom, a detailed non-concordance listing in the lower left text field allows the user to select a particular patient case and then retrieve the rest of the epilepsy record for detailed study.

As mentioned earlier, the GUI was developed using the commercial multimedia package GainMomentum. We are currently developing an XML-based GUI to link to popular statistical packages, such as SPSS and SAS, for more powerful data analysis and exploration. The analytic power of such an image data warehouse increases with the increase in processed patient images and medical cases.

The current approach to lateralization of epileptic foci is qualitative and often depends on the experience of individual imaging specialists. Decision thresholds are numeric thresholds for multimodality lateralization and localization of the epileptogenic zone, which leads to efficacious resective surgery. These thresholds are determined by analyzing the neuroimages of many patients for volume loss in MRI, hypometabolism in PET, and metabolite loss in MRSI.

Many studies have shown a high correlation between the degree of MR-defined volume loss in the hippocampus ipsilateral to the seizure focus and the severity of mesial temporal sclerosis. This threshold is defined in terms of the percentage ratio between the abnormal and normal volumes of identical contralateral brain regions.

Interictal 18FDG-PET usually reveals a single unilateral region of variably mild-to-severe hypometabolism that is used to calculate two abnormality thresholds. One is a ratio with the metabolism of the contralateral region and the other is a ratio with the entire contralateral hemisphere tissue. In proton MRSI, the ratio N-acetyl-aspartate/(chromium + choline) is calculated for identical contralateral regions in both patients and normal control subjects. Patients have decreases in the metabolite concentrations on both sides of the brain, although the decrease on the side of the epileptogenic focus is more severe than on the contralateral side. The image data warehouse provides a platform for investigating the quantitative decision thresholds in epilepsy imaging diagnosis.

**Discussion**

In this paper, we have described a comprehensive, multi-tier image data warehouse framework and the corresponding software development process for supporting neuroimaging research and large-scale data analysis. This modular information framework is based on the OOAD and component-based development paradigm. We believe that the information framework can be generalized for other medical imaging applications. For our initial implementation, the focus is clinical epilepsy. We are steadily extending the scope of the application to include data of other diseases types, such as multiple sclerosis, brain tumors, digital mammography, and lung nodule cancer.

Our experience is that many software components can be reused or customized for different clinical applications. Notable changes in GUIs are needed because different specialties have different user requirements and needs for viewing and querying the data. Needs for image processing algorithms, such as coregistration, quantitation, and correlation, also differ, since most available image processing algorithms are optimized for a particular imaging modality or for particular body organs only. This makes the development of image data warehouse solutions a close collaborative effort among image processing researchers, biomedical informatics scientists, and corresponding disease specialists.

We are currently evaluating such an image data warehouse to enhance users’ practice and to investigate medical hypotheses. We are also gathering user feedback to make iterative improvements and incorporate more data analysis functions for testing hypotheses. Meanwhile, we continue to collect and process various patient disease cases to increase the subject size, data content, and disease types included in the image data warehouse.

Future work will explore the image data warehouse in three application areas. First, we plan to test the scalability of the seizure data warehouse to include data for patients who undergo gamma knife radiosurgery in addition to data for patients who undergo resective surgery. Second, we plan to test the generality of the framework by using it to analyze other neuroimaging-related disease progress and treatment methods, such as those for multiple sclerosis and brain tumors. Third, we plan to extend the level of data abstractions by including genotype information. The genometric approach to studying the genetics of particular diseases attempts to identify phenotypic linkage based on
genotypic linkage. The statistical correlation and data exploration functions provided by data warehouse systems would be an important vehicle in advancing research and understanding in this emerging area of bioinformatics.

References