A Data Warehouse for Integrating Radiologic and Pathologic Data
Daniel L. Rubin, MD, MS\textsuperscript{a,b}, Terry S. Desser, MD\textsuperscript{b}

**Purpose:** Much of the information needed for radiology teaching and research is not in the picture archiving and communication system but distributed in hospital information systems throughout the medical enterprise. Our objective is to describe the design, methodology, and implementation of a data warehouse to integrate and make accessible the types of medical data pertinent to radiology research and teaching, and to encourage implementation of similar approaches throughout the radiologic community.

**Methods:** We identified desiderata of radiology data warehouses and designed and implemented a prototype system (RadBank) to meet these needs. RadBank was built with open-source software tools on a Linux platform with a relational database. We created a text report parsing module that recognizes the structure of radiology reports and makes individual sections available for indexing and search. A database schema was designed to link radiology and pathology reports and to enable users to retrieve cases using flexible queries.

**Results:** Our system contains more than 2 million radiology and pathology reports, and allows full text search by patient history, findings, and diagnosis by radiology and pathology. RadBank has helped radiologists at our institution find teaching cases and identify research cohorts.

**Conclusion:** Data warehouses can provide radiologists access to important clinical information contained in radiology and pathology reports, and supplement the image information in picture archiving and communication system workstations. We believe that data warehouses similar to our system can be implemented in other radiology departments within a reasonable budget to make their vast radiologic-pathologic case material accessible for education and research.

**Key Words:** Databases, data warehouses, radiology reports, data integration


**INTRODUCTION**

Radiologists frequently need to locate interesting cases with particular imaging findings for education or to identify patient cohorts with particular diseases for research studies. Searching for case material is often a difficult and time-consuming task, relying on chart reviews and searches of case logs and individual reports, or working with administrators of medical records or a picture archiving and communication system (PACS). In addition, coverage is limited to the source of information that can vary depending on the hospital database and its purpose in the enterprise.

The current PACS is a rich source of historical image data, but it generally lacks the full scope of patient information needed for research and teaching, such as radiology reports, pathology reports, diagnostic codes, and laboratory data. Even the PACS that contains a radiology information system (RIS) generally does not provide the ability to search for cases that have particular findings or diagnoses, because it is a production clinical system and was not designed to meet the specific needs of radiology teaching and research. Consequently, few radiologists currently have the capability to mine integrated radiology-pathology databases.

Data warehouses are dedicated computer systems that are specialized for integrating related data stored in different source databases with diverging representations or storage formats. Such integrated resources have proven their value in organizing and improving access to medical data, and enabling data-mining activities in the health care enterprise [1,2]. Most existing data warehouse systems focus on the medical enterprise as a whole, without particular attention to the specific information needs of radiologists. We believe that a data warehouse designed with radiology information needs in mind could be use-
ful to radiology education and research. Despite the potential benefits, few data warehouse systems have been implemented in radiology departments to date, possibly because of the technical challenges and costs of creating and maintaining these resources.

Our hypothesis is that it is possible to create an integrated data warehouse for radiology at low cost that can meet many of the information needs of radiology teaching and research. We have developed an architecture for a radiology data warehouse (RadBank) that integrates radiology, pathology, and clinical data into a single searchable resource that we believe can be replicated in other institutions. By limiting the scope of data included to the subset that is most pertinent to radiology, a useful and cost-effective resource can be built and deployed. Such resources may enable radiologists to efficiently identify case material for teaching and retrospective research studies.

METHODS

Regulatory Approval

We obtained institutional review board approval of RadBank as a resource for teaching and research. Waiver of informed consent to include patient data was granted because of the size of RadBank and minimal risk to subjects. We also obtained Health Insurance Portability and Accountability Act approval at our institution, because RadBank needed to contain patient-identifying information to link information from various source clinical information systems.

Separate institutional review board approvals were obtained for individual studies and queries that used RadBank data. Queries performed using RadBank were restricted to those approved under the individual protocols.

Data Warehouse Architecture

The RadBank data warehouse system uses a multitiered architecture that separates the following components: (1) back-end data source systems, (2) data concurrency service, (3) data-staging services, (4) structured report generation, (5) core data warehouse server, and (6) user interface. Figure 1 shows the architecture of the RadBank system. In creating RadBank, we found it operationally helpful to divide the information storage and flow in the medical enterprise into two types: historical data (data acquired before establishing RadBank) and concurrent data (data acquired after deploying RadBank). Our system architecture components handle these two types of information differently to manage the heterogeneity of hospital information systems and diversity of data-storage formats in different systems.

Back-End Data Source Systems. Historical data were extracted from clinical production information systems to populate RadBank with clinical information that was recorded before establishing the data warehouse. We obtained raw data extracts from the RIS (IDXRad, IDX Systems Corp, Burlington, Vermont) and the pathology...
specimen and results database. All radiology reports were extracted using a query to IDXRad in its native language (MUMPS). All available pathology reports were retrieved from that department’s archival database (running Structured Query Language [SQL] Server, Microsoft Corp, Redmond, Washington) with a single SQL query. We also obtained patient demographic data and diagnostic information in the form of International Classification of Disease codes [3] from the hospital enterprise medical information system. These data extracts were in the form of tab-delimited files that could be processed automatically by computer programs without requiring understanding of the underlying storage schemas of the source databases from which they were derived. The data type and legal values of each extracted field were recorded as metadata that were used to design the schema for the RadBank data warehouse. Each record from each database contained all available patient demographic data, which were subsequently used by the data-staging services to prepare the data for RadBank.

**Data Concurrency Service.** To update RadBank with new data as they become available, we created a data concurrency service that reads Health Level Seven (HL7) [4] data streams generated by hospital systems. HL7 is a messaging standard that permits diverse medical information systems to exchange health information. We obtained the HL7 version 2.4 Interface Engine Standard used by our institution and identified those message segments containing the data that RadBank should acquire from the back-end data source systems. Working with the hospital Information Technology group, we created a computer service that collects all HL7 messages pertinent to the data warehouse and forwards them to RadBank by secure file transfer protocol.

**Data-Staging Services.** The data-staging services manipulate, clean, and transform source and concurrent medical data before storing them in RadBank (Figure 1). A core task in data staging is mapping the data schema of the RadBank data warehouse to incoming data streams and reconciling different representations of the same data in different systems (eg, sex may be represented as “M” or “1”). The Historical Data File Extraction module reads the source files containing historical data from the various back-end data source systems and inputs them into RadBank. The HL7 File Extractor parses the concurrent HL7 messages and updates RadBank with the new data. A master field mapping table is used to relate data fields in raw incoming data formats and the fields in the RadBank data schema.

We also created a Patient Identification Reconciliation service that relates patients to their radiology and pathology reports. The medical record number (MRN) is not completely reliable in uniquely identifying patients; a particular patient is sometimes given multiple MRNs, and occasionally the same MRN is associated with more than one patient. We built a master patient index using our institution’s billing and registration database, and we obtained lists of multiple-patient or multiple-MRN assignments from the medical records department. Before a patient is associated with a pathology or radiology report based on the MRN, the Patient Identification Reconciliation service validates the patient identity in the master patient index. If there is a potential patient identification issue, the record is flagged for identification to be resolved manually.

**Structured Report Generation.** To give RadBank the capability of restricting the search to particular sections of radiology and pathology reports, we created a software module that identifies and labels these sections using eXtensible Markup Language (XML) [5], a computer-readable format. This module uses a rule-based approach and regular expression pattern matching [6] to recognize report headings, and it adds the appropriate XML markup to demarcate each section of the report (Figure 2). The rule patterns were
constructed such that sections that are similar but labeled differently will be identified using the same XML element, enabling the computer to recognize them as being the same despite having dissimilar labels. For example, the rule pattern “HISTORY|CLINICAL DATA|CLINICAL HISTORY → <HISTORY></HISTORY>” means that when the module encounters a section of the report labeled with one of the text strings (separated by “|”) specified on the left hand side of the rule, then that block will be marked up using the single XML pattern specified on the right hand side of the rule.

Data Warehouse Server. The RadBank data warehouse was built using a 2.4-GHz single processor personal computer with 1 GB of memory and 2 × 120-GB hard drives running the Linux operating system (RedHat Linux 9, RedHat Inc, Raleigh, North Carolina) and a relational database system (Oracle 9.2.0 database, Oracle Corporation, Redwood Shores, California). The data warehouse has a single database schema that is used to hold, manage, query, and analyze the data in RadBank. This schema is also used to define the interface between RadBank and the various input data sources (historical back-end sources and concurrent HL7). Patient demographics, diagnosis, and radiology and pathology reports are linked using patient identifiers. To improve search performance, indexes were built on selected data fields, such as MRN and patient name, and the text content in individual sections of reports.

The data warehouse layer includes two service components: an XML-parsing component and a data-loading component (Figure 1). The text report and other patient information acquired in the data-staging layer is entered into the data warehouse by the data-loading component. The XML parsing component is used to create an XML marked-up representation of the report to enable queries limited to specific sections of it, such as “Findings” or “Impression” (Figure 2). This layer also includes database and operating system backup and recovery services that archive the system software and database to a separate physical machine on a nightly basis. These were implemented using shell scripts that compress and copy the data, coordinated using an automatic scheduling service provided with the Linux operating system.

User Interface

Currently, the user interface to query RadBank is an SQL query interface. SQL is a database query language that enables a broad range of queries to access databases in flexible ways. We created a set of SQL query templates based on the most frequent search requests to RadBank. There are templates for searching RadBank for keywords occurring in designated sections of radiology reports or pathology reports. There are also templates for searching for radiology reports from particular modalities and for reports that have pathology results within a specified time window of the radiology report. A sample SQL query is shown in Figure 3.

RESULTS

In assembling the initial data for RadBank, historical radiology reports were acquired dating back to 1991 and pathology reports were acquired dating back to 1995. The system currently contains 1,877,596 radiology reports for 351,968 patients and 276,699 pathology reports for 128,563 patients.

RadBank was built at low capital cost ($1,700 for hardware; no cost for public domain software tools) and modest employee cost (one full-time programmer for system construction and maintenance). Consultation with the hospital’s Information Technology group was also required to access legacy data and establish HL7 streams of concurrent data. The design, development,
and deployment of this resource required 1 year. There has been an ongoing cost of one full-time programmer to maintain the RadBank system. The resources required for this system have been reasonable because the component information sources contributing to RadBank are limited in scope and the component architecture of RadBank allows the individual component programs to run autonomously and to be individually updated without needing to reengineer substantial pieces of the overall system.

Since the deployment of RadBank, this resource has been used by many of our radiologists for a variety of teaching and research purposes. To date, it seems to have been a useful resource compared with the alternative of manually searching records. Specifically, we have used RadBank to identify teaching cases demonstrating particular findings that were being sought by radiologists in our department. For example, some of our radiologists searched for teaching cases and located only one case of porcelain gallbladder and two cases of nephrocalcinosis by searching case logs and work lists on the PACS workstation. By comparison, RadBank identified 13 cases of porcelain gallbladder and 23 cases of nephrocalcinosis.

Not all of the cases retrieved by the RadBank system are relevant to user queries; nonrelevant reports are retrieved when the search term appears in a report as a negative phrase (eg, “no evidence for . . .”) or when the term only appears in the history section of the report and the search does not exclude the history section. The total number of cases retrieved by most queries to RadBank is usually limited, however, so false-positives usually are not excessive in the search results.

We have also used RadBank to identify cases needed for an ongoing research study: to find patients who had pelvic ultrasound examinations and were subsequently found to have pathologic evidence for retained products of conception. We found 454 cases in the data warehouse of which 75 had pathologic confirmation, whereas the unassisted radiologist who looked for these cases before the advent of RadBank found 101 cases, none of which had pathologic confirmation. This particular search illustrates the value to radiology research provided by the data warehouse in linking radiology and pathology data. The radiologists who reviewed the RadBank query results reported that the query result format summarizing radiology and pathology results side by side was helpful and simplified the task of identifying relevant cases (Table 1). Furthermore, they reported that even when there were nonrelevant cases in the query output, those cases could be quickly eliminated by scanning the tabular output.

From the queries we have done to date, the research and teaching activities of our radiologists have the following information requirements: (1) identify patients who have particular radiologic findings; (2) retrieve imaging studies with particular radiologic findings; (3) find patients for whom pathology specimens confirm certain diagnoses; (4) find patients who were imaged with a particular

<table>
<thead>
<tr>
<th>Date</th>
<th>Procedure</th>
<th>Radiology Report</th>
<th>Pathology Specimen</th>
<th>Pathology Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>6/10/2002</td>
<td>US pelvis, non-OB, without transvaginal</td>
<td>Intrauterine fetal demise with RPOC</td>
<td>Uterus, intrauterine contents, dilatation and curettage</td>
<td>Products of conception</td>
</tr>
<tr>
<td>6/23/2002</td>
<td>US pelvis, non-OB, with transvaginal</td>
<td>Echogenic material in the endometrial cavity suggestive of RPOC</td>
<td>Uterus, intrauterine contents, dilatation and curettage</td>
<td>Products of conception</td>
</tr>
<tr>
<td>7/31/2002</td>
<td>US pelvis, non-OB, with transvaginal</td>
<td>Findings compatible with embryonic pregnancy with probable RPOC and/or hematoma</td>
<td>Uterus, intrauterine contents, dilatation and curettage</td>
<td>Products of conception</td>
</tr>
<tr>
<td>12/2/2002</td>
<td>US pelvis, non-OB, without transvaginal</td>
<td>Distention of the endometrial canal with fluid and heterogeneous, hypervascular tissue, consistent with RPOC</td>
<td>Uterus, intrauterine contents, surgical procedure not specified</td>
<td>Blood clot</td>
</tr>
</tbody>
</table>

Note: US = ultrasound; OB = obstetric; RPOC = retained products of conception.
Sample output from RadBank query. This table shows the result from the query in Figure 3. The four cases shown here were relevant candidate patients for the research study that prompted this query. RadBank query results show radiology and pathology results side by side, which streamlines the task of identifying relevant cases for research and teaching activities.
imaging modality and who had a biopsy of a particular organ; or (5) find patients with particular imaging results and pathology results. When searching radiology or pathology reports, the radiologists nearly always were interested in restricting their search to particular sections of the report, such as the history, conclusions, or final pathology diagnosis. These requirements could be fulfilled by RadBank because the internal structure of reports are explicitly marked and indexed. The XML structure added to the report permits queries that can be limited to particular sections of reports, such as “history,” “impression,” and “specimen.”

The RadBank queries to search for teaching and research cases were written within a few minutes by the system programmer. SQL query templates were created on the basis of recurring patterns of queries, and these templates reduced the difficulty of composing new queries. Queries to RadBank usually require only a few seconds to run.

DISCUSSION

The amount of patient information available in electronic form within hospitals is exploding, both image and nonimage data. Large repositories of clinical information acquired during routine clinical care provide opportunities for composite analyses that have been exploited to support health care process improvement and epidemiologic research [1,2,7]. Many educational and investigative activities in radiology could likewise benefit from access to the large amount of patient information in electronic hospital resources.

However, in most institutions, it is difficult to search across all data types (eg, radiology and pathology) for a given patient because the clinical, radiologic, and pathologic information is usually scattered throughout the medical enterprise in disparate computer systems, including patient registration and billing systems (health information system), radiology reporting and scheduling systems (RIS), pathology tracking and reporting systems, and PACS. These resources are not designed to support teaching and research activities—they focus on single-patient, transaction-based care rather than data mining across patients. Even in an integrated RIS/PACS, the functionality is generally limited to archiving images and their reports; this system does not integrate the breadth of clinical, radiology, and pathology information. Consequently, it is generally not possible to link information related to the same patient across the various clinical information systems (eg, radiology and pathology).

Data warehousing is a technology that leverages the data contained in existing resources throughout the medical enterprise, allowing researchers and administrators to access and integrate data contained in a diversity of source databases [8]. A related approach is to create software tools enabling “virtual” integration of data distributed within different systems [9,10]. Data warehouse solutions can be expensive and time-consuming to implement and costly to maintain [11], perhaps accounting for the lack of wider adoption of these technologies across institutions.

In our work, we have highlighted that teaching and research activities in radiology have particular information requirements that differ from those of clinical care and that could benefit from a data warehouse such as we have described here. Instead of retrieving all data on a particular patient (a requirement for clinical operations that could be fulfilled by electronic medical record systems), teachers and researchers need to search across many patients to find those matching the criteria of a specific query. They need to identify patients who have particular radiologic findings or particular imaging results and confirmed pathologic diagnoses. They also need to link radiology results to clinical diagnosis. We have shown that a data warehouse such as RadBank can enable radiologists to identify all patients who share similar characteristics, such as having particular radiology findings or pathologically confirmed cases. We have also demonstrated that the cost for creating such a data warehouse is not necessarily prohibitive, although there is some cost for development and maintenance.

One might suggest that it should be possible to use existing hospital information systems or PACS to access patient data for teaching and research. Although there are a few operational systems that can provide data access and query, most hospital systems do not offer this functionality because they are designed solely to meet the information needs of clinical care. Even in those systems in which query is possible, the ability to link across disparate data, such as finding patients with particular radiology results and pathology results, or being able to search within particular sections of reports, is generally not available.

Even in institutions that have existing data warehouses, those systems generally do not meet the specific information needs of radiology, such as identifying patients who have particular radiologic findings or particular imaging results and pathologic diagnoses. Most current data warehouses focus on the specific needs of internal medicine, oncology, and epidemiology [2,8,11,12]—not radiology. Few systems directly link radiology to pathology reports to permit direct query for radiology studies based on pathology results. A query such as “find all cases of retained products of conception that have pathological confirmation” would not be easily performed with many existing systems because several individual searches would be needed (search all radiology reports for pelvic ultrasound cases, and among those, find cases with the suspected diagnosis; then search
the pathology database to determine whether any patients identified had pathologic findings at the time of the ultrasound). Finally, these information resources are generally not in the control of the radiologist, and access is dependent on the availability of those resources and cooperation of those who maintain those systems.

Our goal was to develop a resource to support radiology teaching and research that accomplishes two objectives: (1) integrate those medical data most pertinent to radiology and (2) provide functionality that supports the information query needs of radiology. Our data warehouse brings the subset of patient data most pertinent to radiology together into a single resource. RadBank is similar to other data warehouses in medicine; however, the latter systems were built to address the specific needs of nonradiologic applications, such as infection control, billing and financial analysis, tracking use patterns, and continuous quality improvement [1,2,12-14]. The RadBank data warehouse focuses on combining radiology, pathology, and patient diagnostic information into a single searchable resource. By focusing RadBank on a subset of all patient data, we keep the project streamlined and cost-efficient. We believe other institutions could adopt the methods described and create similar systems.

An important feature of our system is the ability of users to search within specific sections of reports (“section-specific search”). Radiology and pathology reports contain structure—they contain sections describing clinical history, observed findings, conclusions, and final diagnosis. Searching the full text of reports for keywords without regard to structure, such as searching for cases of pneumonia, would pick up reports stating “R/O pneumonia” in the history, as well as those with “pneumonia” in the impression. By targeting the search to particular sections of reports, the specificity of retrieval can potentially be improved. To enable section-specific search, the system uses XML to demarcate the report sections. XML is becoming a standard for structured markup and exchange of medical data, and it is being incorporated into health information standards such as HL7 [15].

Our preliminary experience suggests RadBank may improve access to clinical data for teaching and research, and it may streamline the ability of radiologists to acquire cases for these purposes. RadBank seems to be particularly helpful in finding cases with pathologic confirmation—a difficult task to perform by manual case-by-case search.

There are limitations to our system. First, there are other types of medical information, such as images from a PACS, which are not currently included in RadBank. The schema for the RadBank data warehouse can be extended to add new data types, such as images. It would be possible to link data in RadBank to PACS images by storing a reference to the images with the report data. A second limitation is that we have not yet tested the ability of RadBank to handle growing volumes of data over time. We believe the RadBank design will meet the needs of expanding future data, because it is based on proven data warehouse technology.

Another limitation of RadBank is the SQL-based query interface, which requires queries to be built and executed by a system programmer. We plan on creating a Web front end to RadBank that will automatically create the SQL queries on the back end while providing a graphic interface to allow the user to compose queries. We have found that most queries can be created from “query templates,” so the complexity of creating the SQL query can be hidden from the user. A final limitation is that we have not formally compared RadBank with alternative existing hospital systems to evaluate the incremental benefit of our approach. At our institution, radiologists cannot execute the types of queries presented in this article; until RadBank, the radiologists searched for cases they needed by hand or requested searches from medical records administrators. Since RadBank became available, our radiologists commonly use it in queries for teaching and research.

An alternative to creating the generic radiology data warehouse described here is to develop specialized “boutique” databases or logs to track information about particular types of patients of interest [16-19]. Such resources usually require close oversight and ongoing manual curation to keep them current and complete. Boutique databases are generally useful in their target domains, but they are difficult to scale in a cost-effective manner to include the full volume of clinical studies generated in the medical enterprise. Furthermore, the data-modeling approach in these systems tends to be customized to specific applications, and integrating boutique databases with other hospital information systems is challenging.

We believe our work is applicable to other institutions, and our implementation and results can be duplicated in those settings. The biggest challenge to implementing RadBank in other institutions is the cost of technical personnel to set up and maintain this resource. Because the scope of RadBank focuses on a subset of all the data in the medical enterprise, we believe other institutions can build and maintain a similar resource in a cost-efficient manner. If other institutions develop similar data warehouses, it would be possible to share or exchange summary (de-identified) information.

**CONCLUSIONS**

An integrated data warehouse that supports the information needs of radiology can be created within a reasonable budget using open source tools and conventional database methods. This resource puts the entire radiologic–


pathologic clinical case repository of a hospital at the disposal of educators and researchers.

ACKNOWLEDGMENTS

The authors thank Dr Leslie M. Zatz for assistance in creating the scripts to extract data from the clinical radiology repository; Ethan Lynn for extracting pathology reports from the clinical pathology database; and Terry Moore for assistance with the HL-7 clinical data feed. We also thank Hernan Mejia and Dr Charles Dibble for their assistance in acquiring ongoing clinical diagnostic data from the Stanford Hospital medical information system.

REFERENCES


