A Virtual Repository Approach to Clinical and Utilization Studies: Application in Mammography as Alternative to a National Database

Lucila Ohno-Machado, M.D., M.H.A., Ph.D., Aziz A. Boxwala, M.B.B.S., Ph.D., John Ehresman, Darrell N. Smith, M.D., M.Sc., Robert A. Greenes, M.D., Ph.D.
Decision Systems Group, Brigham and Women's Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115.

A national mammography database was proposed, based on a centralized architecture for collecting, monitoring, and auditing mammography data. We have developed an alternative architecture relying on Internet-based distributed queries to heterogeneous databases. This architecture creates a "virtual repository", or a federated database which is constructed dynamically, for each query and makes use of data available in legacy systems. It allows the construction of custom-tailored databases at individual sites that can serve the dual purposes of providing data (a) to researchers through a common mammography repository and (b) to clinicians and administrators at participating institutions. We implemented this architecture in a prototype system at the Brigham and Women's Hospital to show its feasibility. Common queries are translated dynamically into database-specific queries, and the results are aggregated for immediate display or download by the user. Data reside in two different databases and consist of structured mammography reports, coded per BIRADS Standardized Mammography Lexicon©, as well as pathology results. We prospectively collected data on 213 patients, and showed that our system can perform distributed queries effectively. We also implemented graphical exploratory analysis tools to allow visualization of results. Our findings indicate that the architecture is not only feasible, but also flexible and scaleable, constituting a good alternative to a national mammography database.

INTRODUCTION

Enormous progress in information technology has occurred since the proposal of a national mammography database in 1995[1]. Advances in networking and the Internet, client-server architectures, and database mediators[2] have shed new light on how to build a national resource for mammography information to serve the purposes of auditing, monitoring and tracking patients. In the original proposal, the national mammography database is conceived as a centralized repository of data. Different institutions performing mammography contribute data to the centralized repository on a regular basis. The centralized repository is a physical database residing at a specific institution, as depicted in Figure 1. This approach has advantages and disadvantages[3]. The advantages are (1) optimal query performance, (2) use of a common vocabulary, (3) easy global consistency checking, and (4) simple architecture. However, these advantages are coupled with drawbacks. The main disadvantages of the centralized architecture are (1) need to fit data into a common schema, (2) total dependence on the availability of resources at the central host, (3) need to fit concepts into a common vocabulary, (4) inability to use existing data for direct entry into the central repository, (5) limited feedback to participating researchers, and (6) dependence on a single database vendor and platform.

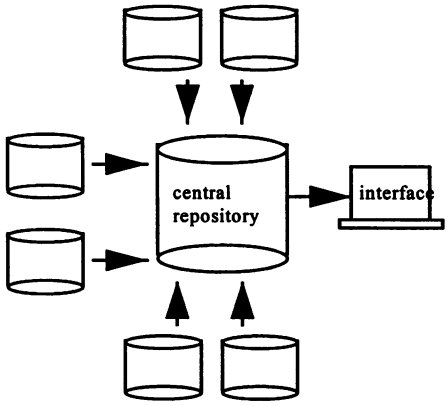

Figure 1. Centralized physical repository. Identical databases at each site send data to central repository. Usually, a single interface is used for queries.

The recent uses of the Internet, and, in particular, of the World Wide Web, to transmit health care data[4], have created new alternatives for the construction of a national source of mammography data. The national mammography database, instead of being a physical centralized repository of data in a single format can be a virtual repository (VR) consisting of a federation of databases at different institutions (Figure 2). When queries are made to the virtual repository, the resulting data are gathered from participating databases. The results of the query can

1091—8280/97/$5.00 © 1997 AMIA, Inc.

either be downloaded for further processing or can have an ephemerous life, lasting only as long as needed to answer a specific question. The databases that participate in this virtual, distributed database federation can not only serve the purpose of contributing data to the virtual repository, but can also be used within an institution in daily practice for clinical and administrative purposes. This approach avoids the redundant entry of data in multiple databases, and allows information to be collected on-line at the point of origin. The addition of exploratory data analysis tools to operate on the results of the queries provides a means for researchers to get timely feedback on their data and to compare their data to those of other centers. For large epidemiological studies, the aggregation of data from multiple institutions has the potential of increasing the power of statistical tests.

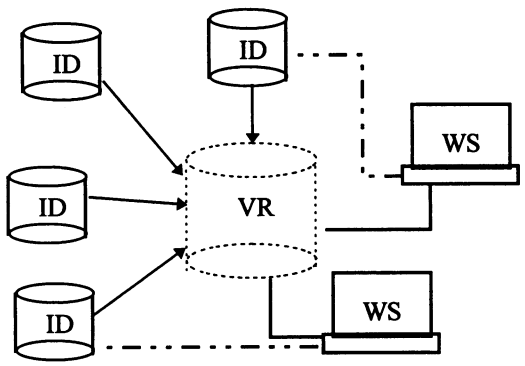

Figure 2. A virtual repository (VR) consisting of a federation of institutional databases (ID). No central physical repository of data exists; when users make queries from their workstations (WS), the results are gathered from the participating databases. The institutional database can be used locally for clinical or administrative purposes.

MATERIALS AND METHODS

We implemented a prototype mammography VR at the Brigham and Women's Hospital to demonstrate its feasibility.

Data

We prospectively collected data from 213 patients who had mammograms interpreted by one of the authors (DNS). These data were collected both for individual tracking of patients by the provider and to be part of our virtual repository. A subset of the data elements and outcome measures proposed for the national mammography database were included in our data set: patient identification number, age, interpretation of mammogram, follow-up recommendation, and results of cytology or biopsy. Encoding of the interpretation of the mammogram and follow-up recommendation was compatible with the BIRADS nomenclature[5], as in the national mammography database proposal. Tables 1 and 2 show the items collected for the interpretation of breast micro-calcifications and masses, respectively.

Table 1. Descriptors of micro-calcifications recorded in the mammography database.

Item	Example values
Distribution	Grouped, linear, diffuse
Size	5 mm or less, 5 to 10mm
Morphology	vascular, skin, coarse, spherical

For both calcifications and masses, the location (clock position) and the laterality were recorded. Additional associated findings such as skin and nipple retraction, skin thickness, skin lesions, axillary lymph node enlargement, architectural distortions, breast density, and follow-up recommendations were also recorded in BIRADS- compliant terminology.

Table 2. Descriptors of masses recorded in the mammography database.

Item	Example values
Margins	circumscribed, obscure, spiculated
Morphology	round, lobular, oval
Density	fat, low, isodense, high

Implementation of V R

The data for all the patients was split into two groups and stored in different databases: an Postgres 95 Server on a HP™ Unix platform and a Microsoft SQL Server™ Version 6.0 on a Windows NT™ platform. Both databases contain information compatible with BIRADS and can be accessed via the Open DataBase Connectivity (ODBC) protocol[6].

The architecture implemented for query formulation is depicted in Figure 3. In this implementation, the user connects to the VR server via a World Wide Web (WWW) browser, and fills a query form similar to the one shown in Figure 4. A Common Gateway Interface (CGI) script is used to generate queries. These "common" queries are translated into specific SQL queries for each database by a mediator, which contains the mapping from the schema the query form uses to each of the database schemata. A database specific query is sent to each database and the results are merged by the mediator.

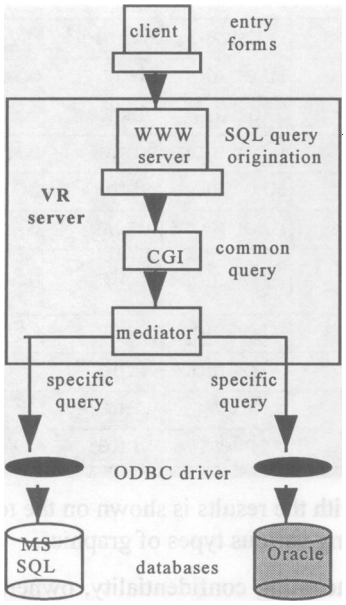

Figure 3. Query formulation system. Common queries are translated into database-specific queries by the VR mediator.

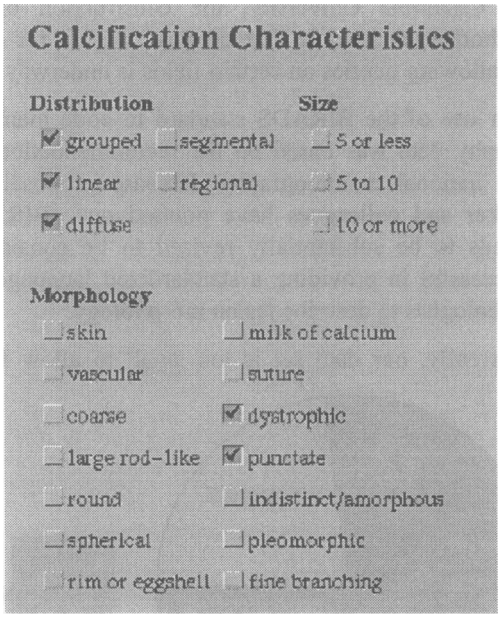

Figure 4. Query form. Checking items in the same category (e.g., Distribution") causes the system to add an "AND" to the query. Checking items in different categories causes the system to add an "OR" to the query.

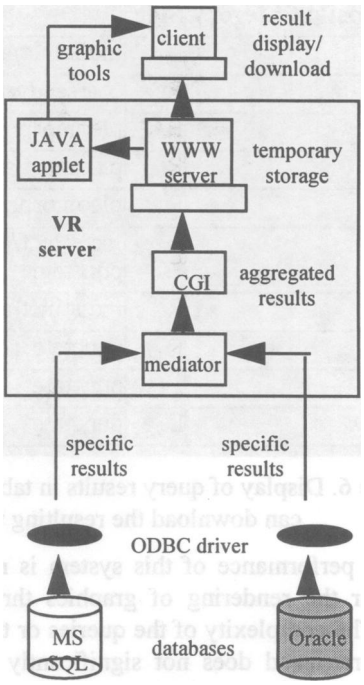

Figure 5. Query retrieval. Results from different databases are aggregated in the VR server, and displayed to the user.

Figure 5 shows the flow of information for query retrieval. The results are aggregated by the mediator into a single table. This table exists temporarily on the VR server, and is removed when the session is over. The user is offered the options of viewing the raw data (Figure 6), downloading the data into a file using various delimiters, or visualizing the data using the graphical exploratory analysis tools such as bar charts, scatter plots, and histograms (Figure 7).

RESULTS

The use of this preliminary prototype has given us an opportunity to get early feedback from users on the value of this approach, and important suggestions on improving the graphical user interface. Some users found the query forms and the selection parameters for the graphical tools difficult to understand.

The radiologist who collected the information uses the system as a personal database of patients. He is currently the only user with access to patient identification. He uses the system's search capabilities, for example, when he wants to retrieve a case of clustered calcifications to use in a teaching session, or to retrieve data from a particular patient. All users are allowed to browse non-identifiable data, download query results, and use the graphical tools.

Diagnostic Category	Side	Morphology	Size	Ethnicity	Age	Site
4	L	indistinct/amorphous, punctate	10 or more	black	65	BWH
4	R	coarse, dystrophic, indistinct/amorphous	5 to 10	black	65	BWH
4	R	pleomorphic, punctate	10 or more	white	55	BWH
4	R	pleomorphic, punctate	10 or more	white	50	BWH
4	R	pleomorphic, punctate	10 or more	black	67	BWH
4	L	indistinct/amorphous, pleomorphic, punctate	10 or more	white	45	BWH
4	R	indistinct/amorphous, punctate	10 or more	white	65	BWH
4	R	punctate, round	10 or more	white	57	BWH
3	L	punctate	5 to 10	white	36	Biocare
3	L	punctate	5 or less	white	40	Spence B

Figure 6. Display of query results in tabular format. A simple table with the results is shown on the top. Users can download the resulting table and/or visualize data using various types of graphics.

The time performance of this system is reasonable, except for the rendering of graphics through Java applets. The complexity of the queries or the number of cases retrieved does not significantly affect the performance time in this prototype, which we believe is more strongly limited by network bandwidth.

DISCUSSION

This work was derived from a collaborative project on building a multi-center virtual repository for breast disease, with participation of Columbia University, University of Chicago, and the Brigham and Women's Hospital. Our long-term goal in that project is to build a virtual repository of breast disease-related data (not only mammography data) that connects existing databases at each institution. We developed the local prototype of the VR to experiment with the architecture. We anticipate that issues concerning confidentiality, ownership of data, and network security will need to be addressed before the proposed distributed architecture can be implemented on a widespread basis. Researchers at the University of Chicago are currently working on clarifying the requirements of Institutional Review Boards for the approval of such virtual repositories. At Columbia University, the construction of an authorization module that creates "views" for user, disallowing queries on certain fields is underway[7].

Our use of the BIRADS standard to code mammography data was based on the recommendations of the national mammography database proposal. As Baker and colleagues have pointed out[8], BIRADS needs to be substantially revised to be considered successful in providing a standardized language for radiologists to describe lesion morphology.

Currently, our data set is too small to allow for a

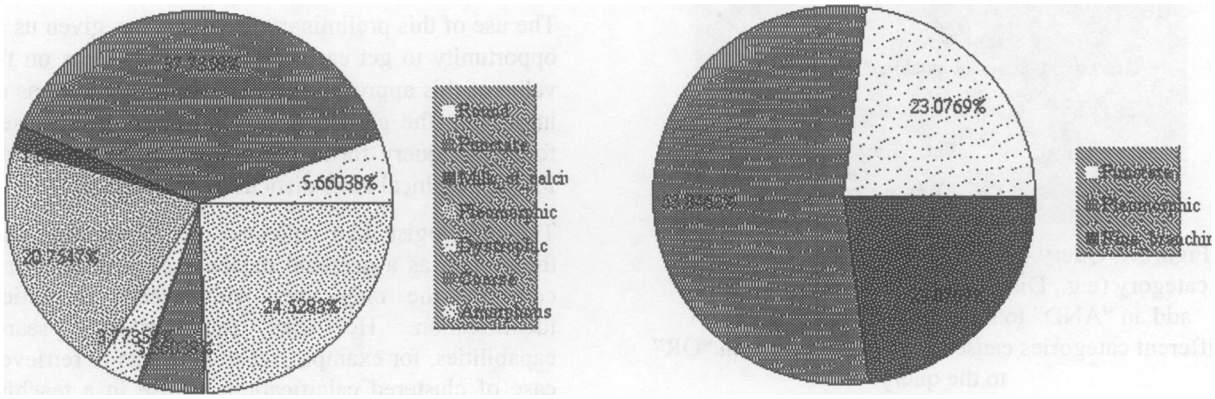

Figure 7. Visualization of query results. In this example, the pie charts show the types of calcification in the mammography report for all patients with a pathologically benign (left) diagnosis and with a pathologically malignant diagnosis (right).

formal evaluation of performance. This preliminary work, however, has allowed us to identify areas in which substantial improvement is needed. As mentioned previously, the user interface for formulating queries is not friendly. We are currently working on an interface that allows users to query by example. The JDBC[9] database driver standard may be substituted for the current ODBC drivers. We are also establishing direct connections between the VR and statistical packages, so that we can possibly bypass the downloading of delimited files and construct graphical interfaces that facilitate the choice of statistical tests. Our current set of exploratory data analysis tools is limited to a small number of graph types. We will expand these graphics with star plots, Andrew plots, and Chernoff faces[10] to facilitate the display of multi-dimensional information.

The translation of common queries into specific queries in our prototype system is done strictly with SQL and a conciliating, non-abstract common schema. We have also worked strictly with two different relational databases, so we have not explored the difficulties in integrating different families of heterogeneous databases. With the trend toward mergers of institutions and the consequent issues of integrating hierarchical, relational and object-oriented databases, sophisticated mediators will be needed to implement the proposed architecture for a national mammography database. Possible solutions for such mediators may include building an abstract data model or query language[11], or creating a common semantic model[12].

We foresee the use of architectures similar to the one proposed here not only in retrospective analysis of cases for auditing practice and tracking patients, but also to prospectively monitor patients in clinical practice or formal clinical trials. The decentralized nature of this system would allow participant institutions in multi-centric clinical trials to use existing databases to provide information for the study coordinator, while keeping control of their local data. The addition of exploratory data analysis tools at all participating sites could contribute to local research at each site.

CONCLUSIONS

We have presented an alternative approach for the creation of a national resource for mammography data. The VR is a decentralized, distributed query system that integrates heterogeneous databases via the Internet. Our prototype implementation suggests that this approach is feasible, flexible, and scaleable, allowing the use of existing databases to populate a national repository.

Acknowledgments

This work was funded by DARPA under contract MDA972-94-3-0047 to the NAMCIC consortium.

References

1. Osuch JR. A proposal for a national mammography database: Content, purpose, and value. *Am. J. Radiology*, 1995; 164:1329-34.
2. Wiederhold G and Genesereth M. The conceptual basis for mediation services. *IEEE Expert*. 1997. In Press.
3. Kim W. *Modern Database Systems: The Object Model, Interoperability, and Beyond.* Addison-Wesley, New York; 1995.
4. Cimino JJ, Socratous SA, and Clayton PD. Internet as clinical information system: Application development using the world-wide web. *JAMIA*. 1995; 2(5): 273-84.
5. American College of Radiology. *Breast Imaging and Reporting Data System (BI-RADS).* American College of Radiology, Reston, VA; 1993.
6. Microsoft Corporation. *Microsoft ODBC 2.0 Programmer's Reference and SDK Guide.* Microsoft Press, Redmond, WA; 1994.
7. Johnson S. Columbia University, New York, NY. Personal communication. 1997.
8. Baker JA, Kornguth PJ, Floyd Jr CE. Breast Imaging Reporting and Data System Standardized Mammography Lexicon: Observer variability in Lesion Description. *Am. J. Radiology*, 1995; 166:773-8.
9. The JDBC database access API. Javasoft Home Page. *http://www.javasoft.com/jdbc/.*
10. Wainer H, Thissen D. Graphical data analysis. *Ann Rev. Psychol*, 1981; 32:191-241.
11. Sujansky W. A *Formal Model for Bridging Heterogeneous Relational Databases in Clinical Medicine.* Ph.D. thesis. Department of Computer Science. Stanford University, 1996.
12. Bressan S, Goh CH, Fynn M, *et al.* Demonstration of the COntext INterchange Mediator Prototype. *Proc ACM ISGMOD/PODS Joint Conf.* 1997. Tucson, AZ.