An Approach to the Modular Design of Medical Research Systems

Justin K. Belcher, Meghan K. Corcoran, Robert J. Han, Hani M. Shageer, Brian A. Welsh, Lisa M. Pastore, and Michael D. Devore

Department of Systems and Information Engineering
University of Virginia
Charlottesville, VA 22903

ABSTRACT

This paper reports on the design and development of a database system capable of collection, storage, and sharing of infertility research data, specifically non-surgical fertility treatments, excluding in vitro fertilization techniques (non-IVF)\(^1\) for researchers at the University of Virginia’s Department of Obstetrics and Gynecology (UVA OBGYN). The created system will be used to study the effectiveness of different fertilization treatment methods, and ultimately improve infertility medical practice. In addition, the design aspects of this system are an important initial step towards the standardization of medical research records, in any hospital and clinic, regardless of medical specialty. The presented system design requirements emphasize database modularity, data distribution, and compliance with HIPAA regulations.

I. INTRODUCTION

Despite the impressive application of information technology in the business and education sectors of the American economy, the healthcare system is far behind the movement towards organizing data into digital format [1]. The organization of information in the healthcare system is highly complex for a number of reasons. First, the amount of data and information being handled has outgrown the scope of in-house information technology applications. Most research and medical organizations do not have the resources to build complex database designs, smooth user interface designs, and other key components of a robust information technology application. Secondly, federal regulations, issues of patient confidentiality, and data security cause the organization of information to be a challenging task.

The design concepts described in this paper are an outgrowth of a project to create an organized system for the collection, storage, statistical analysis, and sharing of fertility treatment research data that would be functional for the UVA OBGYN. Currently, there is an extensive effort in the United States to collect and analyze data regarding in vitro fertilization (IVF) treatments. Federal regulations require that all facilities administering IVF treatments report statistics to the U.S. Department of Health and Human Services Centers for Disease Control and Prevention. However, there is no organized collection of data specifically for non-IVF treatments.

The design addresses many of the problems that arise from integrating an information technology application into the healthcare research field. From a data management perspective, these include simple data entry and management capability, as well as data integrity enforced through a normalized data representation. The technology involved is non-proprietary and widely understood, so the system can be readily extended to support a wide range of research data collections. From a security perspective, these include user authentication and authorization, as well as encryption of sensitive information. Finally, to support collaborative research the system facilitates data sharing by isolating patient identifying data from research data and by permitting a number of distributed system architectures.

A distributed research tool is important because similar research data are collected at different locations to address related research questions. Sharing these data throughout a research community can result in more powerful statistical analyses and can broaden the range of possible research at each institution.

II. DISTRIBUTED COLLABORATIVE RESEARCH

A. Architecture

In order to share data among clinics, the ideal system needs to provide a method to convert the data into a portable form. There are a few different system architectures that support the sharing of data between multiple clinics. One is a hub-and-spoke system (Fig.1). The hub-and-spoke architecture represents a scenario where multiple institutions around the country connect through their web browser to a centralized web server and database through the internet. This design requires the implementation of secure web server and database

\(^1\) Also known as \textit{in vivo fertilization}, an infertility treatment in which conception occurs internally, usually consisting of hormone enhancing drug therapy, distinct from \textit{in vitro fertilization} (IVF) in which conception occurs externally.
connections between the institutions and the central repository. Another potential design would involve a federated database management system (DBMS) (Fig. 2). A federated system integrates multiple autonomous database systems into a single federated database. In addition, this type of system acts as a storage engine that accesses data in tables of remote databases [2]. A federated system should be implemented when multiple offices from one institution want to seamlessly interact with the same data set.

These designs are two extremes, one is fully centralized and the other is fully distributed. When dealing with a data sharing system, a central repository serves as a central location which is responsible for collecting and distributing the data. This ensures consistency in format and updates, whereas a federated database management system may lack consistency in data between institutions. In addition, a centralized data repository acts as a backup of current hard copy data records. However, an organization may choose one method over the other, or a hybrid system depending on organizational requirements.

For the purposes of non-IVF medical research, a hybrid system was created (Fig. 3). This hybrid design includes a local repository at each institution, and the local data collected at each institution is exported to a central repository. From there, the central repository can send out the pool of collected data to each institution. Each clinic could then combine local and imported data for statistical analysis. This design is simpler in that the data is safely transported through a working Import/Export function where information is exchanged periodically. Thus, network access is not required since data can be exchanged by CD or other removable media. As a result, it may be the most secure. The two previously mentioned architectures can be advantageous they provide instantaneous access to remote research data. However, they require all machines to have network access, which require additional security precautions.

**B. Design Requirements**

The designed system can be used as a framework for other medical systems by utilizing the same design strategies. The database covers three areas of medical information: identifiable patient data, diagnoses, and treatments. A class diagram showing the data design for the system is given in the Appendix. All health records are associated with the corresponding patient through a non-informative ID number. With this approach, the research data storage system can be customized to include data from nearly any type of treatment or evaluation through the addition of classes relevant to that specialty. For example, if a researcher wanted to study skin cancer diagnoses, treatments, and results, the tables specific to infertility treatments could be replaced or augmented with tables specific to skin cancer treatments. The overall system structure can remain the same.

To create a satisfactory medical system, the separation of identifiable information is necessary in the design of medical systems. The Health Insurance Portability and Accountability Act’s (HIPAA’s) Privacy Rule forbids the release of this information [3]. In addition, the HIPAA
Security Rule applies to any entity involved in accessing, electronically transmitting, or storing individual identifiable information. Excluded identifiable information includes names, social security numbers, medical records numbers, geographic subdivisions smaller than a state, telephone numbers, dates more specific than the year, or any unique identifying numbers or codes that would distinguish an individual [4].

To accommodate HIPAA regulations, the database features a modular design to isolate identifying information from research-relevant information. This was accomplished by creating a personal information database table that contains all the identifiable data. This table can easily be removed without altering the rest of the database, and none of the other tables are dependent on the personal information table. In all other tables, dates are represented as the number of days since a “day zero” established for each patient. In this way, sequences and spacing between records can be studied without revealing specific dates that could be used to identify a patient.

Complying with HIPAA regulations requires the separation of identifiable and de-identified data. The aforementioned central data repository collects de-identified data from each of the institutions and distributes the collective data set of de-identified data back to various groups. Thus, all of the institutions that are collecting similar data will have an internal data storage system at their respective locations which stores their own identifiable data in addition to the de-identified. Many institutions cannot collect enough data individually to make confident conclusions, but instead, by pooling resources together, they can analyze and test larger data sets and increase the confidence of their results and conclusions.

Another important design requirement of an electronic medical tool is the ability to be freely distributed to various research groups. This can be accomplished by utilizing widely understood and accepted open source technologies, including a relational database for data storage and web interface for data entry. Using these technologies allows the system to be highly distributable, and in addition, encourages data sharing and collaboration.

C. Obstacles

One obstacle in creating a collaborative, robust, medical tool is the lack of medical industry standards regarding data collection. Currently, there are few standards within medical research pertaining to infertility diagnosis and treatment nomenclature. While the International Classification of Diseases, 9th Edition (ICD-9) codes exist, the detail of that coding system is inadequate for research purposes and is restricted to diagnoses only. For example, there are 11 codes in the ICD-9 covering infertility, whereas the current infertility database for UVA OB GYN has 38 infertility diagnosis codes, 16 semen analysis codes, and 32 codes reflecting the female patient’s history of pelvic/abdominal infection or surgery. This poses a problem since the database system at the UVA OB GYN may not integrate well at another research university or organization. Temporarily, a key serves as a guide for researchers unfamiliar with the nomenclatures used in this software system. However, this reduces the efficiency and effectiveness of the project for clinics and organizations unfamiliar with these standards created.

The constant changes within medical research, medical practice, and medical treatments are another obstacle. This system collects data on specified patient information, patient infertility diagnoses, treatments, and results. The project definition narrowed the scope of this system to particular infertility research. Other researchers may want to study new treatments and include new patient information. The ability to add database tables and modify the contents of current tables alleviates this problem.

Additionally, many clinics face the lack of available staff and funding to support the population of the database. In the United States, clinicians may have staff to support insurance and billing services but no data management staff.

III. Security

As any system grows and becomes larger, knowledge of its existence and the data it contains will spread as well. A greater awareness of any protected system increases its risk to be targeted with malicious intent. As this is a natural consequence that cannot be avoided, it is necessary to take precautions protecting the privacy of medical patients.

A. Database Security

One goal of the security measures implemented on the database is to transform individual instances of treatment application into pools of anonymous data for research and analysis. Researchers can use this software to combine resources securely. The database must be secured through methods of user authentication, encryption, de-identification, auditing, and firewalls.

1) User Authentication

A mandatory and preliminary security feature is user authentication. Usernames and passwords are required to access the database, either directly through the database or through the interface developed for the project. Additionally, the computer itself forbids anonymous login, so unauthorized individuals would not even be able to reach the program. Limitations should be set on who can enter, edit, or view data on which tables. Only the administrators will have access to all functions, such as generating new patients and editing current patients’ information, as well as setting access rights to other users. For example, users will only be able to access identifiable information for patients from their own institution.
2) Encryption

All the personally identifiable data are encrypted and password protected in the database. The interface application manages user access so that only authorized users may view and manipulate protected data. Decryption can only occur if the user provides the correct key. Encryption enhances the integrity of the database; however, it slows processing time.

3) De-identification (HIPAA Standards)

Two subsections, or rules, of the Health Insurance Portability and Accountability Act (HIPAA) apply to this system. The Privacy Rule forbids the release of personally identifiable information and the Security Rule applies to any entity involved in accessing, electronically transmitting, or storing individual identifiable information.

The technical security provisions required by the Security Rule of HIPAA require access controls, audit controls, data integrity, person or entity authentication, and transmission security. To be in accordance with HIPAA, one must reasonably and appropriately protect the security and integrity of the researcher’s systems, and the confidentiality of all data displayed, transmitted, or accessed using that system. A layer of confidentiality was designed directly into the system architecture of the database. Distinct tables separate all patient-identifiable information. No clinical data is imbedded in this identifiable information. Once these three tables are dropped, the remaining information in the database becomes anonymous. Researchers can pool clinical data sets and conduct statistical analysis without concern of confidential patient information.

4) Audit Feature

Servers and databases provide audit logs of users and their transactions or of denied access attempts. Actively reviewing logs can identify suspicious activity, such as repeated login failures. These logs capture a variety of information that can be used to track down potential intruders or to alert where security should be reinforced.

5) Firewalls

A secure firewall protection service must be running on the machine. Firewalls provide perimeter protection and separate internal information systems from external. If a database only needs to be accessed locally, Transmission Control Protocol (TCP) networking can be disabled. This means that non-local users will not be able to connect to the database remotely. Enabling this option assembles multi-layered firewall protection.

Additionally for a secure system, the programming language, operating system, and server software must be kept up-to-date to repair and patch any possible exploits. Virus software from a respected security authority must run daily virus scans on the computer, and take the necessary actions when a security threat is present.

B. Secured Interface and Data Transmission

Secure Socket Layer (SSL) technology is used to secure the web-based user interface. This technology provides a private point-to-point transmission of traffic between the client and the server. SSL should be employed whenever sensitive data is available through the Internet. To use SSL for secure communications, a server certificate must be installed and the client must have a browser that can support secure communications. An established certificate authority issues a certificate, which is made up of a public key for cryptographic use, expiration dates, serial number, name, and certificate class. Both the client and the server must have an authorized certificate. The exchange of certificates allows the user to authenticate the sender and then provides for the transfer of encrypted messages [5].

In the situation where the client and server are confined to the same computer, SSL technology is employed to ensure encrypted communication even though data is not transmitted over the Internet. The use of SSL technology allows for the system expansion to include networking between research locations.

IV. IMPLEMENTATION

The main components used to create this system were a web-based interface, an Apache web server, Cascading Style Sheets, PHP, and MySQL, which are all open source technologies so the costs are low. This combination is powerful, flexible, and easily modified. MySQL directly supports data encryption and it also directly supports database federation. The use of web technologies supports the possibility of a pure hub-and-spoke architecture, and does not require any software to be installed on the client machines other than a web browser.

PHP is used to link the database to the user interface. PHP is a server-side scripting language that allows for the dynamic nature of the website. Each time a page is loaded, the web server generates the page content based on real-time queries to the database. In this way the interface will reflect immediate changes to the database to keep the user up-to-date.

In this specific implementation at the UVA OBGYN, a central repository holds all data sent from the individual clinics, and the department’s technical staff restores lost or corrupted data to researchers as needed. The technical staff will run an algorithm that scrubs, or removes, the confidential content in the database before data is distributed between clinics. These individuals are certified to handle the confidential data at this level at the UVA hospital.

V. CONCLUSION

This system produced a database that stores all of the necessary and relevant information for an area of medical research. A web-based user-interface secures a gateway for
direct input of information into the database and allows researchers to query entries in the database. The security measures achieve data confidentiality, prevention of unauthorized access, and data integrity.

This system has satisfied many of the problems that arise from integrating an information technology application into the healthcare system. These problems range from compliance with HIPAA security regulations to the efficient management of health data. The database has a modular design that could be applied to any medical specialty. The data model of the database separates the medical content by topic. Tables containing data on the treatment of cancer or an infectious disease could easily substitute for the infertility treatment tables. The general patient health, personal information, and many other tables would still be applicable. The database design, data entry format, and security measures are applicable to any medical research project.

One limitation of this design type is that it is not very applicable to other fields beyond medical research. For example, it could not be used to study data that requires comparison of exact dates. After the data has been de-identified, the dates are stored as the time period between the initial inquiry and diagnoses or treatment instead of individual dates.

Redundant data entry, a large number of incomplete records, and a limited data pool hindered the efforts of the researchers at the UVA OBGYN. These problems certainly pertain to researchers in many fields of study. This system alleviates many of these issues, and researchers from all fields can adopt these findings to improve the quality and quantity of data available for analysis.

APPENDIX: UML DIAGRAM OF THE DATABASE DESIGN
BIBLIOGRAPHY


AUTHOR BIOGRAPHIES

Justin K. Belcher – is a fourth year Systems and Information Engineering (SIE) student at the University of Virginia.

Meghan K. Corcoran – is a fourth year Systems and Information Engineering (SIE) student at the University of Virginia.

Robert J. Han – is a fourth year Systems and Information Engineering (SIE) student at the University of Virginia.

Hani M. Shageer – is a fourth year Systems and Information Engineering (SIE) student at the University of Virginia.

Brian A. Welsh – is a fourth year Systems and Information Engineering (SIE) student at the University of Virginia.

Lisa M. Pastore – has a Masters and PhD in Epidemiology from the School of Public Health, University of North Carolina at Chapel Hill. Subsequent to a post-doctoral fellowship through the National Cancer Institute and work at an international health consulting firm, she joined the University of Virginia in 2003 as an Assistant Professor of Research. She is jointly appointed by the Departments of Obstetrics & Gynecology and Public Health Sciences.

Michael D. Devore – is an Assistant Professor of Systems Engineering at the University of Virginia. He received his B.S., and M.S., degrees from the University of Missouri-Columbia, and D.Sc. degree from Washington University in St. Louis. He has held a visiting assistant professor in the electrical engineering department at Washington University in St. Louis and prior to that managed system integrity and operations at Amdocs, Inc. He is a member of IEEE and SPIE. He can be reached at mdevore@virginia.edu.