

An Electronic Clinical Research System

B. Timothy Walsh, Jonathan Cohen, and Mandar Patankar

Abstract

Clinical research typically requires the collection of multiple pieces of information from individual research subjects. These data are collected at specific events over a period of time. The nature of the information collected and the schedule according to which the information is collected differ among studies. However, the information gathered in such studies can conveniently be maintained in a relational database. Successful research units typically have more than one study underway at any time, and may complete multiple studies over time.

The goal of the current project is to develop a low-cost, easily maintained, web-based database system (the Electronic Clinical Research System) suited to handle data from clinical studies. The Electronic Clinical Research System permits the design of different studies to be entered, and facilitates the entry and retrieval of clinical data into these studies. Users with different privileges can view, edit, and enter data, and view and edit reports. The system provides the capacity to export data in a format compatible with major statistical software packages, and thereby facilitates the analysis of clinical studies. It is hoped that the Electronic Clinical Research System will be of sufficient value to be used by research groups other than the specific client who requested its development.

Introduction

The broad issue addressed by this project is the systematic storage and retrieval of information collected as an integral part of medical research. In the care and study of individuals participating in medical research, extensive information is collected on each individual. The nature of the information collected varies considerably, but typically includes routine demographic information (e.g., date of birth, sex), the results of staff-administered interviews (e.g., history of previous medical problems and treatment), the results of self-completed forms (e.g., ratings of side effects to a study medication), and results of assessments (e.g., body weight and laboratory tests). In addition, information relevant to the purpose of the study, such as the nature of the assigned treatment (e.g., active medication vs. placebo) must be entered for each subject. In the past, such information was collected on written forms, and specific

parameters were extracted by hand and examined, often using statistical techniques.

Over the last 20 years, as digital information systems have become increasingly available, many of these operations have gradually been transferred to electronic databases. Frequently, however, the research team has little sophistication in database design or database software engines. Therefore, data are often stored in 'flat-file' spreadsheets such as Microsoft Excel [4] which are widely available. However, as the size of the database increases, the lack of a relational structure becomes a major obstacle limiting the users' ability to analyze the data and making large-scale datamining virtually impossible.

More sophisticated investigative teams have begun to employ relational databases. In most instances, these databases have been individually designed for a specific study. In modest-sized and modestly-supported

operations, such databases have been designed and implemented by clinical research personnel, with very limited knowledge of database construction, using PC based software systems, such as Microsoft Access [3]. This leads to a number of potential difficulties. It may be difficult or impossible to combine data from the different study databases, greatly limiting the potential for “datamining” over different studies over different years. The limitations of design and software engines restrict the possibilities of using such databases in a client/server environment and of developing enhancements, such as allowing research subjects to enter information directly onto electronic versions of research forms.

These issues are evident in the database support for the Eating Disorders Research Unit of the New York State Psychiatric Institute at Columbia University Medical Center, which was used as the primary development and testing site for the Electronic Clinical Research System. This research unit was established in 1979, and since that time, has conducted a range of studies of individuals with Anorexia Nervosa, Bulimia Nervosa, and Binge Eating Disorder (a recently identified behavioral syndrome associated with obesity). Currently, there are four major studies underway on Anorexia Nervosa, three on Bulimia Nervosa, and one on Binge Eating Disorder. Research support is derived from the Office of Mental Health of New York State and from grants from the National Institutes of Health and from private foundations. While support is provided for essential infrastructure (e.g., network connections, PC’s), there is very limited support for software engineering to develop software applications appropriate to the needs of the research.

The specific goal of this project was to develop a flexible but stable database system to enable the entry and retrieval of clinical information from research studies underway at the Eating Disorders Clinic, using a

client/server model. The system should require no more than limited support from software engineering and should be usable by staff with little experience with database construction and maintenance.

Relevance

The number of clinical trials underway in the United States is extremely large. For example, the National Institutes of Health (NIH), on its ClinicalTrials.gov Website, lists over 4,000 clinical trials sponsored by NIH. In addition, it is likely that ten times as many studies are ongoing in research centers and clinics around the country without NIH funding. The largest and best funded of these studies will likely have sufficient funding to support IT staff to construct and maintain a study-specific database. However, it is likely that more than 20,000 studies are currently underway which do not have adequate funding or expertise to support the development of a sophisticated study database.

Remarkably, there appears to be no low-cost specialized software available for supporting clinical research. Discussions with a variety of clinical researchers at the Columbia University Medical Center suggests that the following approaches are commonly used.

1. Use of spreadsheets. Spreadsheets are commonly used to record information relating to clinical research. In such systems, data from a single individual is confined to a row and each column contains a single parameter, such as weight or height on a specific occasion.

This approach has several advantages. Software, such as Microsoft Excel [4], is readily available, easy to use, and familiar to most staff. In addition, most statistical software packages, such as SAS [8] and SPSS [9], which are widely used to analyze data from medical research studies, have the facility to import data directly from spreadsheets.

For small studies, with limited numbers of participants and a limited number of parameters, spreadsheets are useful and effective. However, as the number of participants and the number of measurements increase, the use of spreadsheets becomes unwieldy. For example, if a study were to track 100 participants for one year, and to obtain only height, weight, pulse, systolic and diastolic blood pressure weekly, the spreadsheet would have 100 rows and over $5 \times 52 = 260$ columns. One of the studies underway at the Eating Disorders Research Clinic calls for the collection of over 1,000 individual items of data for each subject over the course of one year, and for the enrollment of almost 100 subjects. A spreadsheet with 100,000 cells would be required, and its maintenance and error checking would be extremely difficult. For these and similar reasons, spreadsheets are not ideal for the storage and retrieval of large data sets obtained in medical research studies.

2. Use of Statistical Analysis Software. It is a common practice in medical research studies for investigators to enter data directly into statistical analysis software. Data from medical research studies must be subjected to statistical analysis in order to extract meaningful findings, and, over time, the methods of statistical analysis utilized in medical research have become increasingly sophisticated, typically requiring the use of powerful statistical software packages. Two of the leading statistical packages, SAS and SPSS, typically present data in spreadsheet format. Both packages provide some limited support for data entry, for example, the SPSS Data Entry Station™, but such methods do not permit data filtering, combination, and manipulation which are provided by databases. The ability to collect information across studies, and to flexibly assemble related data is quite limited. For these reasons, the statistical analysis packages have little to offer over spreadsheets in terms of data entry and retrieval.

3. Use of Desktop Database Systems. A somewhat more sophisticated approach to data management is provided by desktop database software systems, such as Microsoft Access [3] and FileMaker [1]. These systems have significant advantages in offering the capacities and strengths of true relational database systems, and are readily available at modest cost. However, several problems arise in adapting such systems to larger clinical research studies. First, such database systems are general-purpose software, not specifically designed to address the needs of medical research studies, which typically call for the collection of similar information on a subject over multiple points in time (e.g., weight over time during an experimental treatment). In order to adapt the desktop database software to meet the needs of a specific study, a significant amount of database design and implementation are required. For example, someone on the clinical research team is required to develop an entity-relationship model and specify the appropriate tables, relationships, and indices. The level of sophistication required is often lacking among the personnel of a medical research team. In addition, the ability to modify the off-the-shelf features of desktop database systems is often limited; modifications in Microsoft Access, for example, must be made in “modules” written in Visual Basic. Furthermore, such systems are best suited for databases of modest size (a few megabytes) on a single PC. They do not scale well to larger studies and are not designed for a client-server environment. Therefore, while such systems satisfy the needs of small clinical research projects, they are less suitable for large projects which may need to manipulate large (several megabytes) of data and provide for simultaneous database access by multiple users.

4. Use of Commercial Clinical Research Systems. A number of companies provide extensive clinical trials management software (e.g., ClinSource [http://www.clinsource.com/] [2], AXIS

[<http://www.axisclinical.com/>] [6], Versal [<http://www.versal.com/>] [10]). These are full-featured services available on contract, and serve the needs of major clients such as the pharmaceutical industry. While pricing was not clear, it is highly likely both that extensive custom modification of these systems would be required to adapt them to the needs of a specific study, and that the fees would be considerable. These systems are intended for studies conducted at multiple sites at multiple locations involving thousands of subjects. They are not available “off-the-shelf” but must be designed specifically for an individual client and project. Such high-end commercial products have a major role to play in well-funded clinical research projects, but are not an effective solution to the needs of modest, not-for-profit research clinics.

In summary, the currently available options for electronic storage and retrieval of clinical research data do not permit an investigator to easily configure a system to maintain and analyze a moderately large clinical research program without the expenditure of substantial funds. The Electronic Clinical Research System (ECRS) was designed to fit this need.

Methodology

The Electronic Clinical Research System is implemented as a web-based system, requiring a database server and networked workstations.

Software utilized. As described above, a review of the available options indicated that no low-cost software options were available to meet the needs of a modest-sized clinical research operation. In this instance, modest sized indicates 2 – 10 clinical research studies, each with 50 – 100 participants, being controlled by less than 5 clinical investigators and a research staff (primarily research assistants and study coordinators) of less than 10 individuals. Because of the severe cost constraints of such environments, it was decided to implement the system using open-source software, if

possible. The only major option uncovered was MySQL [5] as the database server and PHP [7] as the scripting language. These software applications were chosen because they are open-source, thereby eliminating a potentially major cost, they are very widely used, and they are well-supported and well-documented.

Database Design. The initial database design was derived from a ‘legacy’ database developed in Microsoft Access for the Eating Disorders Research Unit. A major driving feature of the design is to develop a database which is both tailored to the specific needs of clinical research but is also completely flexible, so that its methods and relationships can be employed for virtually any clinical research study on any clinical topic.

While the Microsoft Access database was in use for about one year, a number of concerns were expressed by the Eating Disorders Research Unit personnel. Although it was implemented on a client server model, all software was that of Access, which provided both a front-end and a back-end. The transfer of information between the client and server was often slow, and could be interrupted, leading to errors in data entry. In addition, much of the user interface was implemented via Visual Basic, which was poorly documented and extremely hard to modify. Lastly, as the database grew in size, concerns grew that the Access engine would be insufficient to handle multiple simultaneous users and was vulnerable to database corruption. For these reasons, it was decided to employ the basic design but to transfer it to MySQL.

In order to maintain maximum flexibility, the core elements of clinical research studies were reviewed with the client, and the following core elements of all such studies were identified and incorporated into a basic database design.

Study: The name of the study itself.

Factors: Studies typically identify, *a priori*, key factors which will be examined in a study, such as drug versus placebo. These were conceived of as Study Factors, which characterize each subject in the study.

Assessment Points: the occasions at which information is to be collected. Virtually all clinical studies obtain information from participants (patients) at previously identified occasions, called Assessment Points. Each study has a specific sequence of assessment points.

Subject: The participants in the clinical study.

Forms: As is largely true in reality, the data are conceived as being entered on forms, which are completed either by the research subject or the study staff. Each form has one or more questions, and, upon completion, has answers.

Questions: the elements of forms.

Answers: the responses to the questions--the essential data obtained during the study. In order to facilitate the speed of information retrieval, it was decided to maintain all data in a single 'answer' table.

(Currently, this table has approximately 200,000 entries.)

Scores: Many forms are scored, for example, by adding numerical values assigned to answers to the questions on a form. Because these are frequently requested by users, the database contains a table to maintain scores.

Results

The essential elements (Study, Subject, Form, Question, Answer, and Assessment Point) and their relationships are schematically illustrated in Figure 1. This diagram illustrates two crucial features of the design. First, that it is very flexible. For example, though it was designed for an Eating Disorders Research Clinic, there is nothing specific to eating disorders in the design. Second, the key elements are highly inter-related, and there are several many-to-many relationships. Inevitably, this leads to a rather complex database design, especially when additional entities are added (for example, for scores, scoring routines, and study factors) and to implement the several many-to-many relationships.

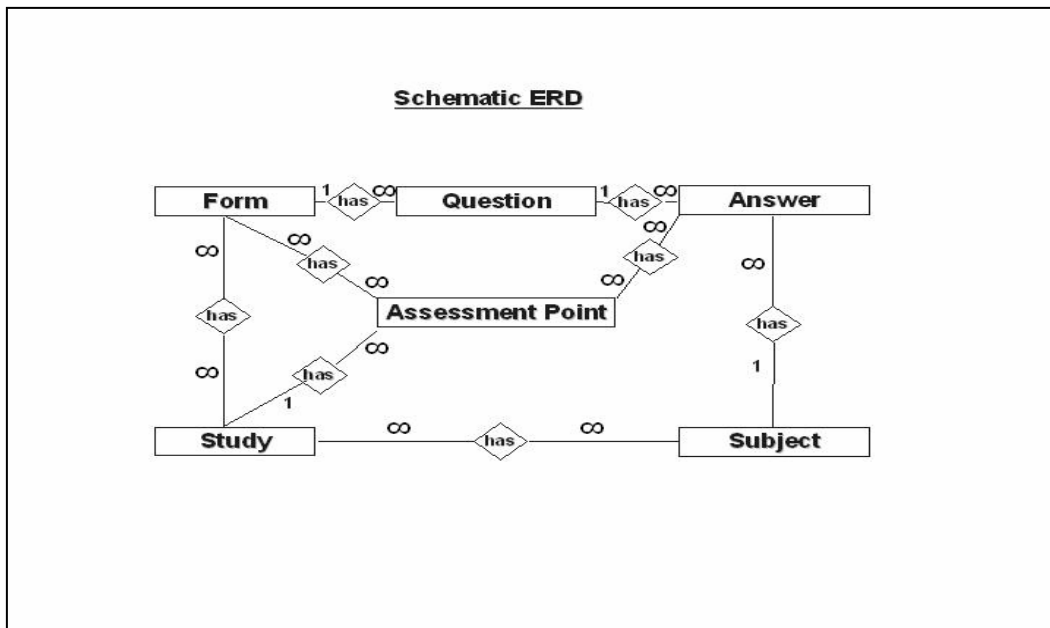


Figure 1: Schematic Illustration of Database

The full implementation of the database contains 23 tables and is shown in Figure 2.

The client interaction with the database has been implemented using php to develop web pages which can be accessed via Internet Explorer. These web pages permit the user to perform all the tasks required for maintenance of a clinical research system without direct interaction with the database itself.

These tasks include:

Study Design:

Addition/editing of a study, including study factors.

Addition/editing of assessment points.

Addition/editing of forms and questions.

Addition/editing of scoring.

Study Maintenance:

Addition/editing of a subject.

Addition/editing of subject data.

Study Reports: Production of user selected data and administrative reports.

The Study Reports function is particularly valuable for clinical research studies. This function allows the investigator to request a report providing specific parameters at particular assessment points for particular subjects. This is typically the core of analysis of the results of the study. The ECRS will allow all this information to be user-specified, and an Excel spreadsheet to be generated with the requested information. This information can then be directly submitted to statistical analysis software for numerical data processing.

Therefore, the database achieves the aims of utility and flexibility, and allows the user to create virtually all the elements required for a clinical research study, to enter and edit data, and to obtain both administrative and research reports.

Discussion

The ECRS represents the development of a software product for an information-processing niche within the broad medical research community. As reviewed above, the currently available products are either too rudimentary and not specifically designed for medical research studies, or custom-built database systems provided by commercial providers at substantial cost. The former are suitable only for very small (single investigator, single research study) programs, and the latter are designed for very large systems, such as the pharmaceutical industry. The ECRS attempts to fill some of the gap between the two.

Several design features of the ECRS are notable. First, it uses client/server architecture. This allows the data to be stored and backed up centrally, and for database operations to be carried out on the server, with limited burdens on information transfer between the user's equipment and the server. The use of this architecture also permits multiple users to access the system from multiple different platforms, requiring only that the user have access to Internet Explorer.

Second, the software (PHP [7], MySQL[5]) used to develop the ECRS is entirely open-source, thereby greatly reducing the cost requirements to the user. In the academic medical research community, IT budgets are typically extremely limited and insufficient to afford commercial database engines, such as Oracle.

Third, the ECRS is designed to be extremely flexible. Although it has been implemented for a specific research clinic dealing with eating disorders, there is nothing in the design which restricts its use to this setting. The key elements of the ECRS are generic to virtually all clinical research studies: the ability to enter information on subjects from forms completed at multiple assessment

points. Furthermore, the ECRS permits the user not only to enter and edit data for specific studies, but also to enter completely new studies and forms without redesign of the software.

Fourth, the ECRS permits the user to obtain reports in Excel spreadsheet format from the database in an extremely flexible and easy to use method. This allows the investigator to transfer information from the study database to statistical analysis systems with great facility.

As of the current writing, the database system has been fully designed and implemented, and the basic data entry and editing php software written. It is anticipated that a fully functional ECRS will be delivered to the Eating Disorders Research Unit by the end of the Spring, 2004 semester, and that on-site testing and actual use will occur during Summer, 2004.

A number of the features of the ECRS would benefit from additional development. Several features of the system were limited to ease development, and might be profitably expanded. For example, the current system has been designed to operate with Internet Explorer. As this is widely available, this does not present a major impediment to use of the ECRS, but it would be best if the ECRS would operate with a range of browsers. The ECRS has been tested on both Unix and Microsoft servers, but testing on other client/server environments would be useful.

A great enhancement would be the implementation of a project developed in CS616 in Spring, 2003. This project allowed subjects to enter data directly into a database, rather than the current system of having the user fill out paper and pencil forms from which information is then input into the ECRS by a research assistant. Direct data entry would save staff time and also reduce transcription errors, and thereby enhance the usability of the ECRS.

Summary

The ECRS is a newly developed flexible system for the entry, maintenance, and retrieval of information derived from medical research clinical studies. The features of the ECRS are not currently available in any off-the-shelf system currently available. If field-testing at the

Eating Disorders Research Unit of Columbia University Medical Center demonstrates that the system functions as designed, it is possible that the ECRS will be attractive to other research investigators. This would be no small accomplishment, and of considerable value to the research community.

References

1. FileMaker Pro7, www.filemaker.com. FileMaker, Inc., April, 2004.
2. Trial XS, 2004. www.clinsource.com. ClinSource. ClinSource UK Branch Office, Derby, UK, April, 2004.
3. Microsoft Access 2002 SP-2. www.microsoft.com. Microsoft Corporation, Redmond, Washington, 2001.
4. Microsoft Excel. 2002 SP-2. www.microsoft.com. Microsoft Corporation, Redmond, Washington, 2001.
5. MySQL 4.1. www.mysql.com. MySQL Inc., Seattle, Washington, April, 2004.
6. *Patient Analysis & Tracking System (PATS®)*. www.axisclinical.com. AXIS Clinical Software, Inc. Portland, Oregon, April, 2004.
7. PHP 4.3.5. www.php.net. April, 2004.
8. SAS 9. www.sas.com SAS Institute, Inc. Cary, NC, April, 2004.
9. SPSS 11.5. www.spss.com. SPSS, Inc. Chicago, Ill, April, 2004.
10. *Versal Metabase*. www.versal.com. Versal Technologies, Inc. Lexington, Massachusetts, April, 2004.