MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Design, Implementation and Management of a Web-Based Data Entry System for *ClinicalTrials.gov*

John E. Gillen, Tony Tse, Nicholas C. Ide, Alexa T. McCray

National Library of Medicine, Bethesda, MD, USA

Abstract

We describe the development and deployment of a web-based authoring capability, the first implementation of which is used for data entry and management in support of the ClinicalTrials.gov web site. The system facilitates efficient collection of summary protocol information from multiple geographically-dispersed organizations. We explain the motivation for developing this capability, and cite critical design goals. We then describe system design, implementation and operation, focusing on essential aspects of each. We conclude with a summary of the extent to which we met our stated objectives.

Keywords:

Software Design, User-Computer Interface, Online Systems, Internet, Clinical Trials, Information Management

Introduction

We have developed a Web-based data entry and management system, the Protocol Registration System (PRS), for sponsors of clinical trials to maintain their records in *ClinicalTrials.gov*, a national clinical trials registry. Section 113 of the Food and Drug Administration Modernization Act (FDAMA), enacted in 1997, requires specific information about all studies of effectiveness for serious or life-threatening diseases or conditions to be represented in *ClinicalTrials.gov* [1]. This information is intended for use by patients and other members of the public, and for health care providers and researchers.

The motivation for creating the PRS came from lessons learned while developing *ClinicalTrials.gov*, first launched in February 2000 [2]. Initially, all data records were submitted to *ClinicalTrials.gov* as XML files through the file transfer protocol (FTP), primarily from Government organizations. Accepting data in this manner proved to be highly error prone, resulting in an iterative, labor intensive error correction process. Furthermore, the anticipated addition of records from hundreds of widely dispersed industry sponsors promised to magnify the problem. To provide a more effective tool for detecting and correcting errors prior to submission, we designed and implemented the PRS.

The PRS facilitates the fulfillment of the requirements promulgated in the Food and Drug Administration's (FDA) Guidance for Industry, published in March 2002 [3]. We collaborated with the FDA to design a web-based system that would not be an undue burden on industry sponsors. In addition, an informational web site containing a guided tour and other information was designed to help familiarize new users with the PRS [4].

The PRS is currently used by more than 300 sponsors to enter and regularly update information pertaining to clinical trial protocols for *ClinicalTrials.gov*. As of February 2004, the site contains approximately 9300 protocol records, representing studies sponsored by the National Institutes of Health (NIH), other Federal agencies, and the pharmaceutical industry, in locations throughout the United States and in some 90 countries worldwide. Protocol record authoring and management is facilitated by a number of validation mechanisms, support tools, internal organizational workflow and control procedures.

Several types of automated authoring tools to assist in the development and management of information have been researched and developed by others. These include implementing narrative clinical practice guidelines as online systems for increased compliance and limiting errors [5], managing specialized biomedical information [6], distributing educational materials [7], and developing and validating clinical trial protocols [8].

Methods

Design Objectives

Our primary design goal was to enable clinical trial sponsors to submit and maintain protocol information in accordance with FDA regulations and guidance [1, 3]. Secondly, it was paramount that the PRS exhibit a high degree of usability in order to minimize the burden on sponsors, thereby encouraging full compliance. Finally, the PRS and associated processes needed to facilitate effective data management and quality assurance, both for sponsors and the PRS staff. Meeting these principal objectives substantially improves the content of the *ClinicalTrials.gov* web site, ultimately benefiting the public.

In addressing our primary design objectives, one factor presenting significant challenges was the fact that the PRS user community is rather widely varied. Sponsors have differing characteristics along a number of dimensions:

- domain and regulatory knowledge
- size: ranging from individuals to large, geographically dispersed organizations
- number and size of protocols, ranging from one single-site study to hundreds of multi-site studies

- computing environments: equipment, software and in-house technical support
- level of quality assurance and independent review

The PRS needs sufficient flexibility to support the full spectrum of sponsors. Organizations with a less stringent level of quality control are of particular concern, as they present a risk of inaccurate, incomplete or outdated information making its way to *ClinicalTrials.gov*.

Another key challenge was striking the appropriate balance between ease of use and obtaining complete, accurate, current and valid data. High usability and reduced burdens on the sponsor tend to encourage compliance with FDA requirements. On the other hand, compelling sponsors to provide the highest quality information better serves the public. For example, prospective participants must have the proper contact information and current recruiting status, if *ClinicalTrials.gov* is to fulfill its fundamental mission.

In order to address the data management and quality assurance objectives, as well as to achieve high usability, a critical consideration was user and administrative workflow. The PRS design had to facilitate effective data entry and review processes, while being flexible enough to support the varied user community. Moreover, to ensure compatibility of our design with sponsors' operations, we had to consider the current workflows of the sponsors at all levels. Workflow was also a key consideration in developing tools for the PRS staff, to facilitate the management of the overall system.

Despite the wide variation among sponsors, it is necessary to provide consistency across all of the protocol records in the system. In particular, the PRS needs to facilitate use of standard terminology, to give consumers the best chance of interpreting protocol information correctly.

It is essential that the PRS exhibit high availability and reliability, as well as consistent performance. This facilitates timely dissemination of clinical trials information to the public.

The PRS needs to be a secure system, meaning that protocol information or information regarding trial sponsors must be protected from unauthorized access. Within the system, one sponsor must not have access to another sponsor's information. Ideally, these objectives need to be met in a manner that minimizes additional burdens on authorized users.

Information Flow

The PRS, in combination with the staff and its processes, addresses the design objectives as illustrated in Figure 1. Clinical trial sponsors access the PRS via any common web browser, entering protocol information and receiving immediate feedback on the validity of that information.

The PRS design and implementation incorporates several categories of information which, either directly or indirectly, affect the processing of protocol data. Validation rules, based on FDA regulations and guidance, are incorporated into the PRS for the validation of user input. Controlled vocabularies are used both for validation and to provide suggestions for alternative terminology or spelling corrections. Data sources used for controlled vocabularies include NLM's Unified Medical Language System (UMLS) for conditions and US Postal Service data for cities, states and countries. User and administrative workflows are reflected in the functionality provided by the PRS and in the details of the PRS user interface design.

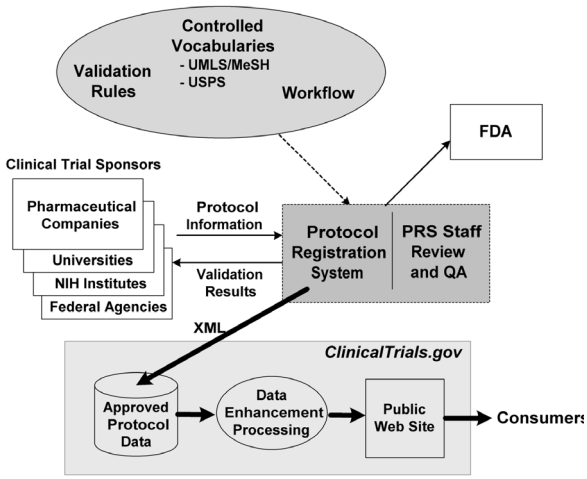

Figure 1 - PRS Information Flow

The PRS staff uses special administrative features of the PRS, as well as other tools, in reviewing protocol information and for releasing approved records to *ClinicalTrials.gov*. Released records are transferred in XML format to the *ClinicalTrials.gov* system for final processing (e.g., adding targeted links to other NIH web sites) in preparation for inclusion on the public web site.

System Architecture

The PRS architecture consists of a network of redundant high-performance servers and components, as shown in Figure 2. The dispatch server directs requests from users' browsers to the PRS web server. The web server hosts the PRS web application, which provides the user interface and the bulk of the functionality of the PRS, including transfer of approved protocol information to *ClinicalTrials.gov*. The web server also hosts the authority services, which implement controlled vocabularies. The database server stores and retrieves protocol information, controlled vocabulary data and metadata upon request by the web server.

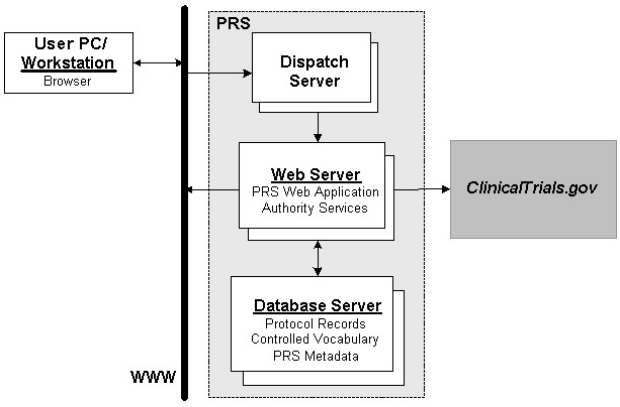

Figure 2 - Protocol Registration System Architecture

The PRS web application is implemented as a Java "servlet" [9]. The selection of servlet technology improves usability significantly, simply by making client requirements minimal. That is, PRS users need not install any special software or browser plug-ins.

The redundant servers in the PRS increase availability and reliability. The dispatch server facilitates automatic failover to a backup server, should the corresponding primary server fail. Two database servers redundantly maintain protocol information, allowing failover without loss of data. Under normal operation, the spare database server takes frequent "snapshots" of the replicated data, allowing for quick recovery, with minimal loss of data, should a system failure occur.

Software Implementation

Performance tends to be a significant issue in web applications, especially given the inherent delays associated with the World Wide Web and many users' internet services. The PRS employs data caching techniques in several areas of the software to minimize the time spent in data retrieval. In cases where data must be retrieved, database queries are optimized using techniques such as indexing.

Adherence to well-established standards improves usability, especially with web-based applications. For example, the PRS generates HTML in accordance with standards which are well-supported on virtually all popular browsers.

Security measures utilized in the PRS include a firewall for the PRS servers, encryption of data transmitted to and from sponsors via HTTPS/SSL, and digital certificates. Sponsors must obtain accounts, and users must provide a username and password to log into the system. Most browsers in common use handle the encryption and digital certificates without any problems; however, we occasionally must provide assistance to first time users with older browser versions that do not support encryption standards.

While the interactive web interface works best for most sponsors, there are some larger organizations that prefer to submit data electronically. An organization might already have a large database of protocol information that needs to be translated in some manner for submission to *ClinicalTrials.gov*. For such organizations, the capability to transmit protocol records in XML format, via HTML upload, is provided. With the upload option, unlike the original FTP approach, the sponsor still gets direct feedback from the PRS, so that errors can be corrected with minimal PRS staff intervention.

User Interface Design

The general topic of web application usability has received much attention in the literature [10]. While the PRS user interface was designed using these well-established usability principles, a few specific aspects are worthy of mention.

An essential characteristic of the PRS user interface design is the distribution of the fairly large set of protocol data elements across many detailed editing screens, rather than using one or a few long, scrolled screens for editing. The detailed screens are organized in a relatively flat hierarchy, with each screen containing functionally related data elements. Data entered by the user is saved as each screen is completed. This approach tends to make entering protocol information less intimidating, as well as eliminating the risk of losing a large amount of data input (e.g., due to a PC or browser failure).

The PRS has a consistent layout for editing screens, an example of which is shown in Figure 3. When creating a new protocol record, the user is presented with the complete series of screens, covering the complete set of data elements associated with a study, in a logical sequence. This approach is similar in concept to the "wizard" technique used in PC software installation and configuration tools, and is thus familiar to most users.

Once all editing has been completed, the user is presented with a main editing screen. This screen shows status information, as well as content, for the protocol record and provides "edit" links which allow the user to return to any specific editing screen. This mode of screen navigation is used in maintaining existing records.

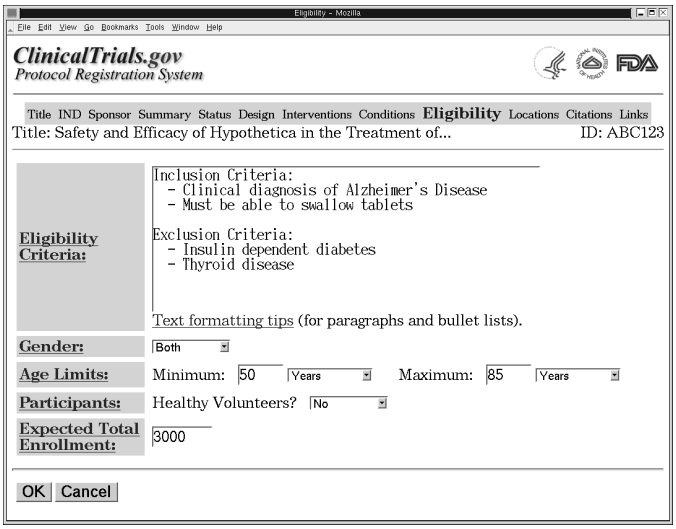

Figure 3 - Editing Screen Example

Data validation is performed when a protocol record is initially selected for display or editing, and screen-by-screen as a record is modified. Validation results are displayed on the main protocol screens, and on the pertinent editing screens. *Error* messages indicate problems that would prevent the record from being published on the *ClinicalTrials.gov* site, such as missing eligibility criteria. *Notes* indicate potential problems, such as an unrecognized sponsor name.

For data elements associated with controlled vocabularies, when an unrecognized term is entered, the user is provided with a list of potential alternative terms. The list is compiled using a combination of spelling correction, lexical variant generation and synonymy from the UMLS.

Data elements consisting of lists, such as trial locations, present a challenge in a web-based user interface, due primarily to the limitations of HTML. The PRS uses a summary list screen with links to an editing screen for a single list item. This greatly simplifies the implementation of functions such as list item deletion, without sacrificing usability dramatically.

To support larger organizations, the PRS provides for both administrator and standard user accounts. This allows one or more administrators to perform functions such as releasing records for publication and monitoring overall organization data entry status, while other users focus on content entry for their assigned protocol records. The PRS also offers an optional feature that allows a very large organization's users to be divided into groups, with different administrator(s) assigned to each group.

For those organizations using both standard user and administrator accounts, the PRS includes features for managing workflow, such as the ability to assign status to records at various stages in the review cycle. The administrator can view a list of records filtered by status and other criteria. The PRS automatically sends email messages to administrators, based on significant events such as a user completing entry of a new record.

Testing

Extensive testing of the PRS, prior to initial deployment and whenever there is a significant upgrade, has resulted in the prevention of software failures and the early exposure of usability and performance issues. For more extensive upgrades we enlist the help of a small group of actual users for testing, which often uncovers problems that were not detected by our development team.

For sponsors who choose to submit information via XML upload, we first provide external access to one of our test systems, allowing them to experiment (and thereby debug their own systems and processes). This allows for a smoother transition when those organizations begin submitting data to the live system, as well as saving a substantial amount of time and effort on the part of the PRS staff.

Training and Support

For new sponsors, the informational web site describes the data required for each protocol record, as well as illustrating that it is not difficult to enter this data using the PRS [4]. The latter objective is accomplished through a series of sample screens that mimic the corresponding PRS editing screens, for prospective users who might initially be concerned about the burden of learning and using the system.

Although we strive to make the system easy enough to use without consulting documentation, the PRS does include an online user's guide. When a usability issue arises that cannot be sufficiently addressed through modification of the screen design, our preferred approach is to place "just-in-time" help on the pertinent screen.

When administrators log into the PRS, or upon request, all of the organization's records are checked for potential problems, such as records that have never been released for publication. If any such problems are detected, the administrator is alerted and provided with a link to generate a more detailed report.

Discussion

PRS utilization is significant, as indicated by the statistics provided in Table 1.

Most sponsors are able to request accounts and commence using the PRS to submit protocol information without any intervention from the PRS staff. Table 2 shows a breakdown of support requests received monthly by the PRS staff, via email, for the year 2003.

Table 1: PRS Usage Statistics

Usage Metric	2003 Monthly Average
Number of user sessions	1687
Protocol records created	452
New or modified records sent to *ClinicalTrials.gov*	1993

Table 2: PRS Administration Support Requests

Type of Request	2003 Monthly Average
Account registration inquiries	8
How-to questions	10
Password reset requests	4
Site access problems	0.5
XML upload questions	1

Certain aspects of the day-to-day operation of the PRS have significant bearing on meeting our design objectives. For instance, by monitoring log files generated by the software, our development team can gain insight into issues such as which features are most (or least) useful or where performance needs to be improved. Even more important, in some cases, problems which would have gone unreported have been detected through review of the log files.

Given the variance in quality assurance practices of the sponsors, it is necessary for the PRS staff to provide final review of all records submitted. Significant time and effort is saved by making minor modifications, such as spelling corrections, ourselves. We also make more significant content modifications in certain cases, such as editing the "conditions" field to use MeSH terminology for better search results. For other content issues, we communicate directly with the sponsor and request that they make the necessary changes. Table 3 shows statistics on the ClinicalTrials.gov protocol records modified by the PRS QA staff.

Table 3: PRS Quality Assurance Statistics

QA Modification Category	Number of Protocol Records*
No modifications needed	6568 (83%)
Minor modifications	766 (10%)
Significant modifications	578 (7%)
Total published records	7912

* As of January 2004

To facilitate system operation and oversight, the PRS provides the capability to generate several types of reports, drawing on metadata maintained by the PRS as well as protocol information. The reports cover topics such as statistics by organization, protocol records recently released by sponsors (for publication on *ClinicalTrials.gov*) and records potentially needing attention (e.g., not updated for too long a time period). Where appropri-

ate, the reports are made available to sponsors, as well as PRS and FDA staff members.

Conclusions

The original *ClinicalTrials.gov* system's reliance on FTP transfer of XML formatted protocol records for input suffered from simple formatting errors, data content problems, and a labor intensive feedback loop. The PRS system solves the formatting problems, since most users enter their data directly into the PRS. The PRS addresses the data content issues by providing option menus for applicable data items and providing error checking for other controlled fields. The situation is even improved in the case where sponsors upload XML files into the PRS, as the user can see formatting errors immediately, without any involvement of the PRS staff.

The PRS provides a work flow that allows the staff to review and perform a quality assurance step on every record before it is released to *ClinicalTrials.gov*. This final step proved to be essential for ensuring that protocol information provided to the public is of the highest quality. With the FTP transfer mechanism this approach would not be practical, as any corrections made by the QA staff would have been lost with the next file transfer.

The emphasis on usability in developing the PRS has resulted in the vast majority of new sponsors being able to get started using the PRS without any assistance from the staff. This has led to better compliance and more timely submission of data on the part of trial sponsors.

Much of the PRS architecture, design and software has been reused or adapted for use in other systems, either deployed or under development, at the NLM. For example: the Genetics Home Reference web site's Content Manager [11]. Similarly, the PRS operational procedures and overall approach serve as a guide for the deployment and support of these systems.

References

[1] FDA Modernization Act of 1997, Public Law 105-115, 105th Congress. Section 113, Information Program on Clinical Trials for Serious or Life-threatening Diseases. Food and Drug Administration web site. Available at: http://www.fda.gov/cder/guidance/105-115.htm

[2] McCray AT, Ide NC. Design and Implementation of a National Clinical Trials Registry. *J Am Med Inform Assoc* 2000; 7(3):313-23.

[3] Guidance for Industry Information Program on Clinical Trials for Serious or Life-Threatening Diseases and Conditions. FDA Web site. Available at: http://www.fda.gov/cder/guidance/4856fnl.htm

[4] Protocol Registration System Information web site. Available at: http://prsinfo.clinicaltrials.gov

[5] Peleg M, Patel VL, Snow V, Tu S, Mottur-Pilson C, Shortliffe EH, Greenes RA. Support for Guideline Development through Error Classification and Constraint Checking. *Proc AMIA Symp* 2002:607-11.

[6] Jakobovits R, Brinkley JF, Rosse C, Weinberger E. Enabling clinicians, researchers, and educators to build custom web-based biomedical information systems. *Proc AMIA Symp* 2001:279-83.

[7] Lehmann HP, Nguyen B, Freedman J. Delivering labeled teaching images over the Web. *Proc AMIA Symp* 1998:418-22.

[8] Rubin DL, Gennari J, Musen MA. Knowledge representation and tool support for critiquing clinical trial protocols. *Proc AMIA Symp* 2000:724-8.

[9] Apache Jakarta Project, Tomcat web site. Available at: http://jakarta.apache.org/tomcat/index.html

[10] Nielsen J. *Designing Web Usability*. Indianapolis: New Riders, 2000.

[11] Genetics Home Reference web site. Available at:http://ghr.nlm.nih.gov

Address for Correspondence

Alexa T. McCray

Lister Hill National Center for Biomedical Communications

U.S. National Library of Medicine

8600 Rockville Pike

Bethesda, MD 20894