Use of a Secure Internet Web Site for Collaborative Medical Research

W. Wesley Marshall, MD
Robert W. Haley, MD

In 1997, we undertook a large clinical research project in which collaborating scientists in 15 laboratories and clinics in different buildings on 3 campuses in 2 cities needed to enter, edit, and verify clinical and laboratory data for patients and controls under investigator blinding.1-4 All investigators continuously generated test results and needed access to core demographic and scheduling information that constantly changed. Central statistical and epidemiologic staff needed access to current data to perform interim analyses for grant extension applications while maintaining the blinding of investigators who were still examining and testing new patients. The investigators in each laboratory needed to remain blind to the information being collected by those in the other laboratories and clinics.

While preparing to set up an expensive system of dedicated telephone lines, high-speed modems, a telecommunications staff, a system for handling and storing paper forms, and a security system to safeguard the project data, we considered numerous alternatives used in recent years5 and decided to pursue considering numerous alternatives used to handle and store data, and a security system to safeguard the project data, we considered numerous alternatives used in recent years5 and decided to pursue

Researchers who collaborate on clinical research studies from diffuse locations need a convenient, inexpensive, secure way to record and manage data. The Internet, with its World Wide Web, provides a vast network that enables researchers with diverse types of computers and operating systems anywhere in the world to log data through a common interface. Development of a Web site for scientific data collection can be organized into 10 steps, including planning the scientific database, choosing a database management software system, setting up database tables for each collaborator’s variables, developing the Web site’s screen layout, choosing a middleware software system to tie the database software to the Web site interface, embedding data editing and calculation routines, setting up the database on the central server computer, obtaining a unique Internet address and name for the Web site, applying security measures to the site, and training staff who enter data. Ensuring the security of an Internet database requires limiting the number of people who have access to the server, setting up the server on a standalone computer, requiring user-name and password authentication for server and Web site access, installing a firewall computer to prevent break-ins and block bogus information from reaching the server, verifying the identity of the server and client computers with certification from a certificate authority, encrypting information sent between server and client computers to avoid eavesdropping, establishing audit trails to record all accesses into the Web site, and educating Web site users about security techniques. When these measures are carefully undertaken, in our experience, information for scientific studies can be collected and maintained on Internet databases more efficiently and securely than through conventional systems of paper records protected by filing cabinets and locked doors.

In developing a secure Web site for research, we found no single article or book that enumerated all the required systems and components, explained how to assemble them to create the secure Web site, or even defined the many new technical terms brought into use because of the Internet. We therefore undertook a thorough literature search of MEDLINE, the Internet, and computer bookstores and consulted extensively with experts in commercial Internet Web site development. Ultimately, we created a secure Web site, used it successfully for 3 years to collect data for our multisite collaborative clinical case-control study, and have since developed secure Web sites for 2 more research projects. In this
article, we describe the practical issues researchers need to understand to establish a highly efficient, secure Internet Web site for collecting and managing research data.

THE INTERNET

In the 1960s, Paul Baran, an electrical engineer at RAND Corp, described the concept of the Internet in a series of technical reports developed for the US Air Force. His ideas included linking heterogeneous computer systems and diverse networks into 1 large network. In 1968, the Advanced Research Projects Agency Network was created.6,7 The introduction of the transmission control protocol/Internet protocol (TCP/IP) in the early 1970s allowed all types of computers and operating systems to be linked to 1 large network.8 This network allowed for the expansion of computer science research by providing remote access to distant computers, remote file sharing, and computer resource sharing.9 Additional important innovations that led to today’s Internet include the World Wide Web (the Web), file transfer protocol, Internet relay chat, telnet, gopher, and e-mail.6,10 The Web, developed by the European Organization for Nuclear Research in Geneva, Switzerland, emerged as the most important of these. According to its principal architect, Berners-Lee, and colleagues,11 the “World Wide Web was developed to be a pool of human knowledge, which should allow collaborators in remote sites to share their ideas and all aspects of a common project.”

The Web uses a client/server model for exchanging information (Figure 1).12 A person who wishes to provide information to others sets up and maintains a server computer that hosts an organized base of information known as a Web site. Others wishing to access information on server computers install software called a browser on their computers, making them client computers. Using the network communication protocols hypertext transfer protocol (HTTP) and TCP/IP, the client browser sends requests to a specific server located potentially anywhere on the Internet.8 After receiving a request, the server computer responds by sending, or serving, the information back to the client computer in the form of images known as Web pages.13 Developed with the powerful computer language called hypertext markup language (HTML), Web pages combine text, menus, bitmap images, and other user-interface elements into a multimedia display for easy exchange of complex information.14 After receiving the HTML-encoded message over the Internet, a client computer reconstructs the elements that make up the Web page and displays them on the computer screen. Because all information published on the Web uses the same programming language, any computer that contains browser software to decode HTML can display the Web page. The Web page itself can contain HTTP links (hyperlinks or hypertext links) to other Web pages on the same server computer or to Web pages on other remote server computers, regardless of the type or brand of the computers hosting them.15

To support millions of Web sites on the Internet, the communications protocol TCP/IP sets up addresses for server and client computers on the In-
Setting up a Web site for managing clinical research data over the Internet has recently become feasible through the development of fast computers and powerful Web browser, server, and security software systems specifically tailored to provide easy, secure exchange of information over the Internet. These components can be assembled to form a powerful research Web site by carrying out the following 10 steps (TABLE 1).

(1) Plan and develop the scientific database that will store the research data. All collaborators in the study create a list of variables that they plan to collect from the various tests proposed. For each variable collected, the collaborator lists the data types, valid code lists (for categorical variables), valid ranges of values for scalar or continuous variables, and the formulas for fields to be calculated from the data they enter (eg, SDs, ratios, indexes).

(2) Choose a database management software system for storing the data. This requires estimating how many investigators will be accessing the Web-enabled database simultaneously. Any of several off-the-shelf relational database management software products are adequate for a study that will have fewer than approximately 30 users accessing the database at the same time. If the database will be accessed by more than 30 investigators at one time, an enterprise-level database is needed.

(3) Create tables in the database to store the information. The first table will contain information that identifies the study participants. It usually contains each participant’s name, medical record number, study number, contact information, and medical history data. Subsequent tables will contain fields for the research data to be entered by each investigator. Allocation of a separate table in the database for each investigator will simplify the setup for keeping each investigator blind to the data entered by other investigators, if required by the study design, and will make it easier to export the investigators’ data files for their own analysis. Each investigator’s data tables can then be linked via the patient study number or medical record number to a subject-identifying table.

(4) Develop the screen layout for the Web site, typically using an HTML editing software program. Each data entry item on the Web page screen relates directly to a field in a database table. The screens are typically organized in a hierarchical system that allows researchers more efficient access to their data entry screens. The hierarchical screen layout can be used as part of the security system to confine researchers’ access to their own data entry screens, if required.

(5) Choose a software product called middleware that connects the database management system to the pages of the Web site. Middleware contains the instructions that tell the Web browser on the client computer how to locate each field in the database on the server computer. Middleware can be thought of as a set of “virtual wires” that connect each data entry item on the client computer screen with its corresponding field in the database on the central server computer (FIGURE 2).

(6) Embed in the Web pages the computer code that checks the accuracy of the data entered. These rapidly executed software routines, developed using JavaScript scripting language, allow range and code checks, data correction, and field computations to be carried out on the client computer before the data are transmitted to the central server database. With JavaScript routines embedded in the Web pages, the client’s computer becomes an interactive Web page with rapid response time and powerful, easy-to-use data entry and error-checking capabilities.

(7) Set up the server computer that will host the Web site in the central data management office and load the middleware and the database management system that will maintain the data. In choosing an operating system for the server computer, important considerations include sufficient built-in security measures to ensure confidentiality, availability of advanced middleware software tools, and ability to handle processing requests by multiple researchers simultaneously. The Web site manager installs the database management software system and the middleware system on the server computer to begin hosting of the Web site.

(8) Obtain an official, unique TCP/IP Internet address and a distinct DNS name for the new Web site by contacting the Internet Network Information Center (InterNIC; http://www.internic.net), a cooperative entity composed of the National Science Foundation and...
Figure 2. Middleware Connecting Fields in Database Tables With Data Entry Items on Web Page

Table 1 in this example database contains input variables and Table 2 contains derived variables calculated by JavaScript code that is embedded in the Web page.

Table 2. Security Measures to Protect a Scientific Web Site

<table>
<thead>
<tr>
<th>Protecting Client Computers</th>
<th>Protecting Server From Internal Corruption</th>
<th>Protecting Server From External Invasion</th>
<th>Preventing Interception of Information Exchanged Over the Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Use an operating system with process isolation (eg, Windows NT/2000, not Windows 95/98) to prevent interference by invading terminate-and-stay-resident programs and viruses</td>
<td>• Locate the server in a locked room</td>
<td>• Install a firewall device between the server and the Internet to prevent bogus information packets from reaching the server computer</td>
<td>• Use encryption on information packets transferred between server and clients</td>
</tr>
<tr>
<td>• Obtain client certification for each client computer from a certificate authority</td>
<td>• Protect the server’s power with an uninterruptible power supply and surge protection</td>
<td>• Avoid naming the root login page to prevent search engines from discovering the Web site</td>
<td>• Use domain restriction to exclude access from computers with unapproved transmission control protocol/Internet protocol addresses</td>
</tr>
<tr>
<td>• Authenticate users by authorized user name and password or by biometrics (eg, fingerprint, voice scan, retinal scan)</td>
<td>• Limit the number of people who have access to the server computer</td>
<td>• Require a valid user name–password combination to allow access to the Web site</td>
<td>• Set up audit trails to record dates and times of all user access to the server</td>
</tr>
<tr>
<td>• Use timed logouts to prevent unsupervised access to the system</td>
<td>• Set up the server computer as a stand-alone computer, not connected to any other network</td>
<td>• Give the uniform resource locator name only to the investigators participating in the study</td>
<td>• Prevent users from saving or bookmarking Web pages below the login page</td>
</tr>
<tr>
<td>• Install virus detection software on the server computer</td>
<td>• Require additional user authentication for access to the server computer and database</td>
<td>• Disable all forms of Internet communications on the server computer, including e-mail, telnet, file transfer protocol, Internet relay channel, hyperterminal, and gopher</td>
<td>• Educate the Web site users on security techniques</td>
</tr>
<tr>
<td>• Install a firewall device between the server and the Internet to prevent bogus information packets from reaching the server computer</td>
<td>• Prevent interference by invading terminate-and-stay-resident programs and viruses</td>
<td>• Use domain restriction to exclude access from computers with unapproved transmission control protocol/Internet protocol addresses</td>
<td>• Use encryption on information packets transferred between server and clients</td>
</tr>
</tbody>
</table>

INTERNET SECURITY

Security is one of the primary concerns in collecting medical research data with a database system that can be accessed over the Internet. Most clinical research on patients involves confidential information that must not be accessible by anyone outside the research staff. Use of the Internet is a 2-way exchange of data over a public information network. Before data are exchanged, each message is divided into equal-sized units to which leading address labels and trailing termination markers are attached to form information packets.

Just as the Internet makes it easy for a server computer to exchange information with its users on client computers, it also makes it possible for computer hackers to break into the server computer and steal the information, corrupt the system by sending bogus messages to it, or intercept information packets as they stream between the client and server computers. Such incidents may be acts of random vandalism, or they may be attempted simply for the challenge, with the intention of stealing the actual data, or with deeper intelligence intentions (eg, embedding programs to attack industrial or government computers).

The security of a Web site has been aptly compared with that of a medieval castle. The Web site developer, like a medieval king, must protect the central databases from internal corruption and external invasion and must safeguard against interception of communications between the central server and remote client computers. Specific precautions must be taken to protect the client computers, the server computer, and their Internet communications (Table 2). Two relatively recent advances that have made high-level security possible are firewall technology and public key cryptography.
FIREWALL TECHNOLOGY
A serious threat to a research Web site is that ill-intentioned computer hackers may try to send packets of information into the server computer to harm the server computer and database (eg, a Trojan Horse attack) or disrupt client-server communications (denial of service attacks).20,33 Computer hardware and software programs called firewalls have been developed to restrict bogus packets and applications from reaching the server computer.33 Firewalls may provide 4 types of security functions, including packet filtering, application gateway, circuit-level gateway, and proxy functions.22,34 A firewall usually resides on a dedicated computer or other hardware device, called a firewall device, that is placed between the server computer and its connection to the Internet (Figure 1). By using at least 2 of the 4 types of security functions, the firewall recognizes authorized (“friendly”) information packets and distinguishes them from unauthorized (“enemy”) packets by verifying the types of information packets, the frequency of packet passing, and whether requests for packets are made by computers with a valid address in the security network.22

PUBLIC KEY CRYPTOGRAPHY
A common technique used by ill-intentioned hackers is to intercept packets of information being sent back and forth between a server and its client computers (“packet sniffing”).20 A hacker may tap into a communications line, intercepting packets of transmitted information and assembling them to eavesdrop on the communications. In secure Internet Web sites, messages are routinely protected from packet sniffing by the process of cryptography (Figure 3).20,31 Recent advances provide high levels of protection from unauthorized disclosure. As of August 2000, there have been no publicized instances of compromise of encrypted Web communication by network eavesdropping (written communication, Lincoln D. Stein, August 8, 2000).

High-level encryption requires the server and client computers to have a direct security relationship, which is accomplished by a process known as the public key cryptography system.22,29,31 This system involves using the encryption software built into most Internet browsers as well as a unique digital security certificate installed on the server computer and on each client computer. Digital security certificates are purchased from a commercial certificate authority.29,31,35 In issuing security certificates, the certificate authority conducts security investigations of each participating organization and individual at whatever investigatory level the purchasing organization requests. Background checks might be as simple as verifying faculty or staff status at a university or as comprehensive as full security and criminal record checks for commercial or military organizations. On completion of background checks, the certificate authority issues an electronic digital security certificate, consisting of a unique, long numerical string of numbers, for each server and client computer.

The digital security certificates that reside on server and client computers allow each computer to protect (encrypt) and decipher (decrypt) messages using a key-pair system29,31 (Figure 3). When the security system is established, each computer, including the server computer and all client computers, is given a unique key pair consisting of a public key (a code number that can be publicly known) and a private key (a different code number that stays hidden in its disk storage space). The private key is required to decrypt a message encrypted with the public key. To send a secure message, the server looks up the recipient computer’s public key and uses it to encrypt the message. Since decryption of the encrypted message requires the recipient computer’s private key, the recipient computer is the only computer that can decrypt the message. The en-

©2000 American Medical Association. All rights reserved.

(Reprinted) JAMA, October 11, 2000—Vol 284, No. 14 1847
Table 3. An Example of Itemized Costs of Developing and Maintaining a Secure Web Site for Research*

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Cost, $</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firewall device</td>
<td>1000</td>
</tr>
<tr>
<td>Server computer</td>
<td>1000</td>
</tr>
<tr>
<td>Uninterruptible power supply</td>
<td>200</td>
</tr>
<tr>
<td>Operating system</td>
<td>800</td>
</tr>
<tr>
<td>Database</td>
<td>400</td>
</tr>
<tr>
<td>Middleware</td>
<td>1300</td>
</tr>
<tr>
<td>Virus detection software</td>
<td>500</td>
</tr>
<tr>
<td>Web page editing software</td>
<td>100</td>
</tr>
<tr>
<td>Client/server certificates</td>
<td>500</td>
</tr>
<tr>
<td>Operating system</td>
<td>800</td>
</tr>
<tr>
<td>Database</td>
<td>400</td>
</tr>
<tr>
<td>Middleware</td>
<td>1300</td>
</tr>
<tr>
<td>Virus detection software</td>
<td>500</td>
</tr>
<tr>
<td>Web page editing software</td>
<td>100</td>
</tr>
<tr>
<td>Client/server certificates</td>
<td>500</td>
</tr>
</tbody>
</table>

**Developmental Costs**

| Web page design (hypertext markup language/JavaScript) | 2000 |
| Database design and construction                     | 3000 |
| Middleware development                                 | 5000 |
| Security implementation                                | 2000 |

**Yearly Maintenance Costs**

| Daily backups                                        | 1000 |
| Database and Web site modifications†                | 1000 |
| Renewal of client/server certificates                | 200  |
| **Total**                                            | 2100 |

*These costs are presented as examples based on our experience. Costs may vary in other settings and for other programs and functions.
†More complex data structures, errant or changing design specifications, and developer inexperience can escalate costs substantially.

To break the code and decipher the information, the level of encryption is determined by the encryption method and the number of digits used in the key pair: the more digits used, the higher the level of encryption. The degree of difficulty of deciphering encrypted information (d) is an exponential function of the number of bits (b) used in the encryption process, expressed as $d = 2^b$. For example, by the usual trial-and-error method of deciphering (the “brute force” method), deciphering a message protected by 2-bit encryption would require 2^2, or 4, trials to break the code; a message protected by 4-bit encryption would require 2^4, or 16, trials to break the code; and so on. Forty-bit encryption requires 2^40, or 1099511627776, trials to break the code and decipher the information.

The degree of difficulty of deciphering encrypted information (d) is an exponential function of the number of bits (b) used in the encryption process, expressed as $d = 2^b$. For example, by the usual trial-and-error method of deciphering (the “brute force” method), deciphering a message protected by 2-bit encryption would require $2^2$, or 4, trials to break the code; a message protected by 4-bit encryption would require $2^4$, or 16, trials to break the code; and so on. Forty-bit encryption requires $2^{40}$, or 1099511627776, trials to break the code and decipher a single packet. With the ever-increasing computational power of computers, however, the number of bits used in encryption and the complexity of the encryption algorithm will continue to increase. When deciding what level of encryption to use, a Web site designer should refer to the most current standards for the level of security required by the types of information and the level of threat.

**COST OF ESTABLISHING A SECURE WEB SITE**

In our experience, establishing a secure Internet Web site for a research project costs between $20000 and $35000 to implement initially and approximately $2500 to maintain annually, depending on the complexity of the database structure, the completeness of the initial database design, and the level of experience of the Web site developer (Table 3). In other settings, these costs may vary substantially because of differences in experience in designing and managing research projects, programmer salary levels and proficiency, and level of security required.

**CLINICAL RESEARCHERS AND THE INTERNET**

A central activity in clinical research is the collection, management, and analysis of data. In small research projects, data can easily be collected on paper forms and analyzed by hand, but, as the size of the project or the number of collaborators increases, the need to rely on computerization also grows. In recent decades the number of large, complex projects involving collaboration between many investigators and institutions has increased rapidly. To make computers useful, software for database management and statistical analyses has evolved to a high degree of sophistication. The costs of desktop computers has declined as their data storage and processing power has increased. Consequently, powerful and useful computing is now readily available to the clinical researcher.

The most problematic step in automating clinical research has been the ability to tie together the computer systems of multiple collaborators in a clinical research project. Until recently, such computer networks were managed either by mailing paper forms or diskettes of spreadsheet data to a central statistical center for manual data entry or by teleprocessing over leased telephone lines, with high-speed modems, dedicated terminals, and a database management staff maintaining the connections, which were often costly.

In the past several years, the Internet has emerged as a new platform for collaborative data networks. Whereas its use for medical training, educating, and counseling is increasing rapidly, it has not yet been widely used for clinical research. This is primarily because of the perceived inability to ensure the confidentiality of sensitive data on research participants stored in databases accessible by the Internet. Confidentiality is not generally required for training, educating, and counseling applications but is mandatory for research systems. Although high-level security systems for Internet databases have recently become available and are widely used by commercial organizations, the lack of clear descriptions of these systems has limited their use in clinical research.

Once we indentified and understood the diverse sources and created our secure
Web site for collecting clinical research data, we found it to be far superior to previous networked data collection systems. Our colleagues see the same user interface whether they are working on Apple MacIntosh, Unix-based, or IBM-compatible computers running Microsoft Windows 95/98/NT/2000, Apple OS, Unix, or IBM OS/2. Our research colleagues experience rapid response when loading Web pages that display their data entry screens and receive instant feedback from data editing routines and calculated fields that run on their client computers even before being transmitted to the central server computer. Since the client/server model requires only personal computers with network interface cards rather than modems and dedicated telephone lines, a Web site on the Internet is far less expensive to set up and maintain and is extremely stable, rarely never going down. Our Web site manager can monitor and provide service to the Web site from any computer connected to the Internet and can be notified automatically by e-mail or pager when problems arise. Since the central database is instantly updated by collaborating scientists as they enter their data, statisticians and epidemiologists have the most current data available for statistical analysis, linked directly into the database via ODBC (object database connectivity). In addition, such Web-site databases should be able to be accessed with equal speed and response by collaborating scientists anywhere in the world with a connection to the Internet.

Alternative data collection models, although still useful in the right context, are less satisfactory than a secure Web site in the broad range of research contexts. For example, the commonly used system of adopting a common database package and mailing or e-mailing spreadsheet files to a central location may work well for small databases but is difficult to manage for large or hierarchical clinical databases with many data collection sites and collaborators. Using e-mail to send data files is a highly insecure form of data transmission, and use of regular or overnight mail is relatively slow and subject to misrouting and loss.

model consequently requires meticulous accounting by a central database manager to ensure that all sent spreadsheets are received and merged into the database, the most current values for each field are maintained, and observations are not merged more than once. This model can cause frustration when the results of many tests become available sporadically over long periods. Omitting personal identifiers reduces the consequences of security breaches but does not eliminate the security burden because of the ever-present risk of statistical identification of individuals by unique combinations of data.

In our experience, a secure Web site solves many of these problems by placing access to the database in virtually every collaborator’s laboratory. Like the levels of security of a medieval castle, the Web site developer can ensure many layers of protection against security threats from within and outside the organization. By using current security strategies, scientific information collected on Internet databases can now be maintained as securely as computer databases kept on a computer in an organization. By using current security strategies, scientific information collected on Internet databases can now be maintained as securely as computer databases kept on a computer in an organization. By using current security strategies, scientific information collected on Internet databases can now be maintained as securely as computer databases kept on a computer in an organization. By using current security strategies, scientific information collected on Internet databases can now be maintained as securely as computer databases kept on a computer in an organization.

Acknowledgment: We acknowledge David E. Jackson, MD, who assisted with design of the database and the Web site security systems.

REFERENCES


©2000 American Medical Association. All rights reserved.