Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support

Paul A. Harris [a,*], Robert Taylor [b], Robert Thielke [c], Jonathon Payne [d], Nathaniel Gonzalez [e], Jose G. Conde [e]

[a] Department of Biomedical Informatics, Vanderbilt University, 2525 West End Avenue, Suite 674, Nashville, TN 37212, USA
[b] Office of Research Informatics, Vanderbilt University, 2525 West End Avenue, Suite 600, Nashville, TN 37212, USA
[c] General Clinical Research Center, Medical College of Wisconsin, 9200 West Wisconsin Avenue, Milwaukee, WI 53226, USA
[d] Biomedical Research Education and Training, Vanderbilt University, 340 Light Hall, Nashville, TN 37232, USA
[e] Center for Information Architecture in Research, University of Puerto Rico, P.O. Box 365067, San Juan, PR 00936, USA

ABSTRACT

Research electronic data capture (REDCap) is a novel workflow methodology and software solution designed for rapid development and deployment of electronic data capture tools to support clinical and translational research. We present: (1) a brief description of the REDCap metadata-driven software toolset; (2) detail concerning the capture and use of study-related metadata from scientific research teams; (3) measures of impact for REDCap; (4) details concerning a consortium network of domestic and international institutions collaborating on the project; and (5) strengths and limitations of the RED-Cap system. REDCap is currently supporting 286 translational research projects in a growing collaborative network including 27 active partner institutions.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

The R01 funding mechanism may be the cornerstone of America's biomedical research program, but individual scientists often require informatics and other multidisciplinary team expertise that cannot easily be obtained or developed in the independent research environment [1]. The National Center for Research Resources has stated that the future of biomedical research will involve collaborations by many scientists in diverse locations linked through high-speed computer networks that enable submission, analysis, and sharing of data [2]. However, the need to collect and share data in a secure manner with numerous collaborators across academic departments or even institutions is a formidable challenge. This manuscript presents a metadata-driven software application and novel metadata-gathering workflow used to successfully support translational research projects in the academic research environment. REDCap (Research Electronic Data Capture) was initially developed and deployed at Vanderbilt University, but now has collaborative support from a wide consortium of domestic and international partners.

2. Methods

The REDCap project was developed to provide scientific research teams intuitive and reusable tools for collecting, storing and disseminating project-specific clinical and translational research data. The following key features were identified as critical components for supporting research projects: (1) collaborative access to data across academic departments and institutions; (2) user authentication and role-based security; (3) intuitive electronic case report forms (CRFs); (4) real-time data validation, integrity checks and other mechanisms for ensuring data quality (e.g. double-data entry options); (5) data attribution and audit capabilities; (6) protocol document storage and sharing; (7) central data storage and backups; (8) data export functions for common statistical packages; and (9) data import functions to facilitate bulk import of data from other systems. Given the quantity and diversity of research projects within academic medical centers, we also determined two additional critical features for the REDCap project: (10) a software generation cycle sufficiently fast to accommodate multiple concurrent projects without the need for custom project-specific programming; and (11) a model capable of meeting disparate data collection needs of projects across a wide array of scientific disciplines.

REDCap accomplishes key functions through use of a single study metadata table referenced by presentation-level operational modules. Based on this abstracted programming model, studies are developed in an efficient manner with little resource investment beyond the creation of a single data dictionary. The concept of metadata-driven application development is well established, so we realized early in the project that the critical factor for success would lie in creating a simple workflow methodology allowing research teams to autonomously develop study-related metadata in

an efficient manner [3–5]. In the following sections, we describe the workflow process developed for REDCap metadata creation and provide a brief overview of the user interface and underlying architecture.

2.1. Study creation workflow

Fig. 1 provides a schematic representation of the workflow methodology for building a REDCap database for an individual study. The process begins with a request from the research team to the Informatics Core (IC) for database services. A meeting is scheduled between the research team and an IC representative for a formal REDCap demonstration. Key program features (intuitive user interface, data security model, distributed work environment, data validation procedures, statistical package-ready export features, and audit trail logging) are stressed during the initial meeting in addition to providing project-specific data collection strategy consultation. Researchers are given a Microsoft Excel spreadsheet file providing detailed instructions for completing required metadata information (e.g. field name, end-user label, data type, data range, etc) about each measurement in each case report form. They are asked to spend time over the next several days working with the spreadsheet template to define data elements for their specific project, and then return the worksheet to the IC via electronic mail. The returned worksheet is used to build and populate the study-specific database tables feeding a working web-based electronic data collection (EDC) application for the project. A web link to the prototype application is given to the research team along with instructions for testing and further iteration of the metadata spreadsheet until the study data collection plan is complete. The system is then placed in production status for study data collection. The workflow process typically includes several members of the research group and allows the entire team to devise and test every aspect of study data collection requirements before study initiation.

2.2. User interface

The REDCap user interface provides an intuitive method to securely and accurately input data relating to research studies. Fig. 2 shows a typical CRF view. Each form is accessible only to users who have sufficient access privileges set by the research team. Forms contain field-specific validation code sufficient to ensure strong data integrity. In addition to checking for mandatory field type (e.g. numeric entry for systolic blood pressure), embedded functions also check data ranges (e.g. 70–180 mmHg) and alert the end-user whenever entered data violates specified limits. Data fields may be populated using text fields or through embedded pull-down boxes or radio buttons where the end-user is shown one value and a separate code is stored in the database for later statistical analysis (e.g. 0 = No, 1 = Yes).

Clickable project menu items are shown on the right side of the screen in Fig. 2. All menu items in the Data Entry tab point to CRFs specific to the scientific project, while the Applications tab contains menu items pointing to general REDCap functions. Researchers use the "Data Export Tool" to push collected data out of the REDCap system for external analysis and may select entire forms and/or individual fields for export. The module returns downloadable raw data files (comma-delimited format) along with syntax script files used to automatically import data and all context information (data labels, coded variable information) into common statistical packages (SPSS, R/S+, SAS, Stata). The "Data Import Tool" module allows bulk upload of data from existing files with automatic validation of data and audit trail creation similar to those created when using CRF data entry methods. The "Data Comparison Tool" module provides a mechanism to view and reconcile data for those studies employing double-data entry or blinded-data entry methods. The "Data Logging" module provides users a view of all data transactions for the duration of the project. The "File Repository" module allows end-users to store individual supporting files for the project and retrieve wherever and whenever necessary. The "Data Dictionary" module allows researchers to download a clean copy of the project metadata during the iterative study data planning process and the "User Rights" module is used by project administrators to add or remove research personnel and set individual security rights.

2.3. Architecture

The REDCap project uses PHP + JavaScript programming languages and a MySQL database engine for data storage and manipulation [6,7]. Hardware and software requirements are modest and
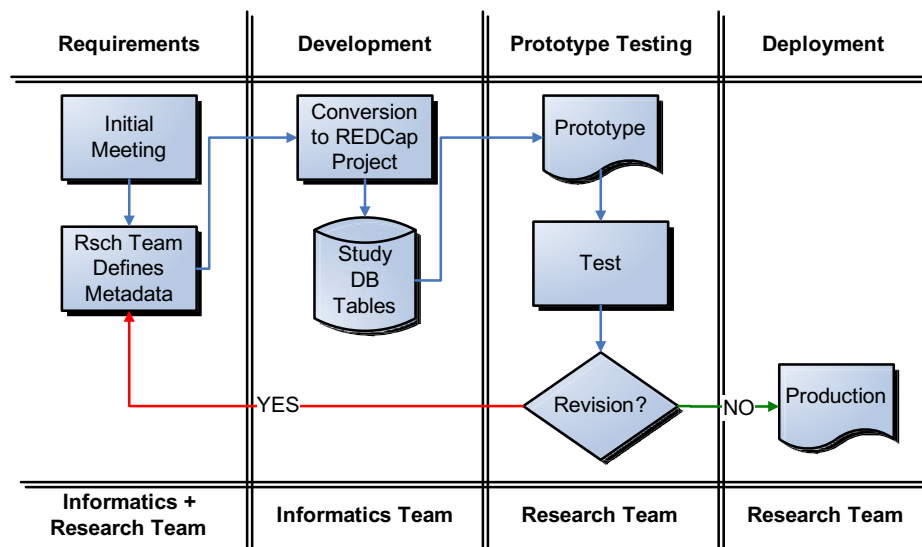

Fig. 1. REDCap Project Initiation Workflow: Study-specific databases are created using data dictionaries provided by the research team. After an initial demonstration, research teams use a custom MS-Excel file template to provide project metadata. The informatics team uses this file to create a prototype web application that researchers can test while revising their data dictionary. Once consensus is reached by the research team on the entire data collection CRF package, the application is moved to production status for study initiation.

Fig. 2. REDCap Data Entry: Case report forms are accessible to users who have sufficient access rights and contain field-specific client-side validation code sufficient to ensure data integrity. In addition to checking for mandatory field type (e.g. numeric entry for systolic blood pressure), validity functions also check data ranges (e.g. 70–180 mmHg) and alert the end-user whenever entered data violates specified limits. CRF pages and REDCap functional modules are accessible to end users by clicking links on the right-side application menu of each project.

the system runs in Windows/IIS and Linux/Apache web server environments. In keeping with the goal of creating a rapid development methodology and easily maintainable resource for multiple concurrent studies, we devised a set of similar database tables for use in each study. The standard REDCap project requires five distinct tables stored in a single MySQL database: (1) a METADATA table containing all study data pertaining to database storage (data field types and naming used for automatic creation of separate data storage table) and CRF presentation (form names and security levels, field-specific display and validation rules); (2) a LOGGING table used to store all information about data changes and exports; (3) a DOCS table used to store uploaded (ex. consent forms, analysis code) or generated export files (SAS, SPSS, R, Stata, Excel); (4) a RIGHTS table containing specific researcher rights and expiration settings; and (5) a flat DATA table used to store all collected data for the study (one row per subject with all collected data fields stored in columns). In studies requiring greater than 1500 data fields per subject, multiple DATA tables are used with a 1:1 relationship between tables linked on the subject identifier field. Although simplistic, this data model is easy to reproduce and tailor for individual research studies during the project creation process and has proven adequate for a wide variety of clinical and translational research studies seen in multiple academic research environments.

3. Results

3.1. Research utilization

The first REDCap project was placed in production at Vanderbilt University in August, 2004. REDCap was successfully deployed and gained a wide reputation within the institution as a toolset and workflow for providing secure, web-based data collection services. Informatics investment has varied during the time of operation, but has typically not exceeded one person performing new module development work, supporting investigators and working with project consortium partners.

Fig. 3 shows activity for the REDCap project at Vanderbilt University and our CTSA partner institution, Meharry Medical College, from August, 2004 through September, 2008. REDCap was initially created to support clinical research studies, but we now also use the project to support basic science and operational databases for our CTSA organization (e.g. process database for managing REDCap projects, accounting activity for CTSA Community Engagement core). The total number of projects is currently 204 (clinical = 189; basic = 7; operational = 8), including 156 active projects and 48 projects in prototype (pre-production) status. The project creation process may take from less than a day to over a year to move from
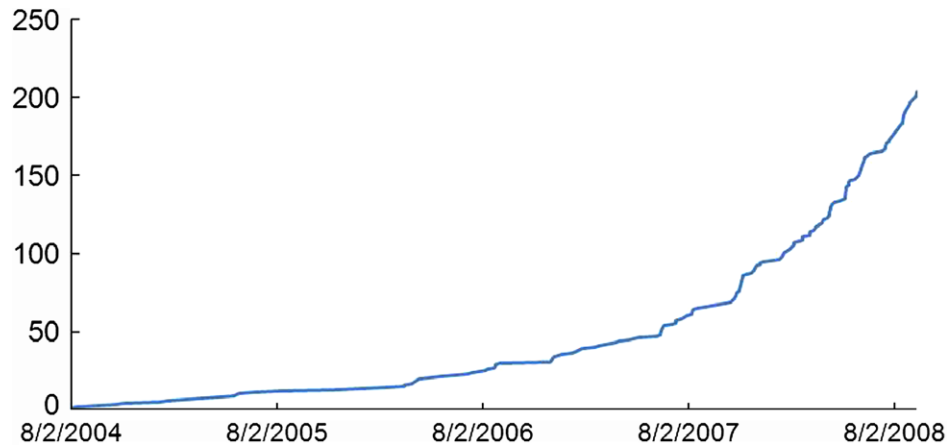
Fig. 3. REDCap Study Utilization at Vanderbilt and Meharry Medical College: Local use of the REDCap project has grown steadily since introduction in August, 2004 at Vanderbilt University and our CTSA partner Meharry Medical College. September 11, 2008 statistics indicated 204 total projects, including 156 active and 48 in the prototype development process. Projects are allowed to remain in prototype status as long as researchers are interested in continuing development.

prototype to production status and depends entirely on the motivation and readiness of the research team in defining CRF requirements and launching the study. Projects are allowed to remain in prototype status as long as researchers are interested in continuing development.

The total subjects in all Vanderbilt and Meharry production-level clinical research study databases currently numbers 17,959 and demonstrates active use for the REDCap system in the clinical domain. The number of registered end-users for all clinical, basic and operational projects is currently 722 unique individuals, including 363 users known through audit logs to have actively participated in data entry or export. An analysis of the number of registered users per project conveys information about the size and diversity of supported studies (median = 4, mean = 5.1, range = 1–77). An analysis of the number of projects per registered user conveys information about the diversity of supported users (median = 1, mean = 1.5, range = 1–14).

3.2. Multi-institutional consortium: software availability and collaborative network

In August, 2006, we launched a pilot initiative to study the process surrounding developing and sharing research informatics resources across academic institutions (Fig. 4). The REDCap consortium now consists of 27 partners (26 domestic, 1 international) collaborating on new development and providing support for REDCap operations. Using this consortium approach, we have successfully added new REDCap modules, created documentation and support processes for adopting centers, added language abstraction (English, Spanish and Japanese versions currently available) and also regularly define and refine near- and long-term plans for improvement and added functionality. We provide software and support at no charge to institutions, but ask for contributions to the overall project through member participation in weekly all-hands meetings and other initiatives. We are encouraged by progress to date and encourage partnership with other academic institutions who would like to join the project. Information concerning joining the consortium may be found at the consortium public website (http://www.project-redcap.org).

4. Discussion

Several factors contribute to REDCap's successful implementation in the clinical and translational research enterprise. Most important, the need for researcher-controlled data services and secure data collection, storage and export is a universal need for any single- or multi-center research study. The initial REDCap demonstration meeting provides an opportunity for project-specific informatics consultation concerning data collection and storage, data validation procedures, data security and logging requirements, and forms layout. The prototype testing and refinement workflow process ensures a team-based data definition strategy prior to study initiation, thereby improving the timeliness and overall quality of study data. The startup time required to launch a new project is short and almost entirely determined by researcher input in defining a structured study-specific data collection plan. After the initial researcher demonstration meeting, IC support for a single project typically requires less than 60 min over the life of a study (including allowances for mid-study CRF modifications when necessary). The Vanderbilt, University Office of Research Informatics offers REDCap services at no charge for scientific research teams and considers the modest personnel investment (<0.5 FTE for universal training and support) a sound investment ensuring researchers have ready access to centralized resources compliant with HIPAA best practices.

A number of limitations should be recognized with the REDCap project. The flat-table structure used for collected research data is inefficient from a data storage standpoint, but very efficient from the perspective of study setup and ideal for embedded data export procedures. The researcher-led metadata creation process is fast and flexible, but does not encourage the use of data naming standards, with the exception that data dictionary templates are typically reused by researcher teams across studies. We are considering methods for adding standards-based data identifiers to the metadata process using IC expert guidance in studies where this approach would add value.

5. Conclusion

The cost of purchasing and supporting major vendor solutions for clinical data management systems can be prohibitive [8]. Although costs can be justified in large organizations running multiple sponsored clinical trials, the academic research environment is typically home to many preclinical investigator-initiated research studies requiring a fraction of the subjects needed for larger trials. In these cases, traditional EDC setup time for CRFs and software expense may outweigh benefits for the study [9]. The REDCap project provides a rapid-development and flexible informatics systems-based approach to supporting studies in the translational research enterprise. The project benefits from a

Fig. 4. REDCap Consortium: Geographical representation of 27 active domestic and international collaborating institutions. Locations depicted do not include sites where software is used, but hosted at another institution (example Venezuela, Argentina, Peru, Mexico, Honduras, Haiti, Chile, and Brazil data entry partner sites).

consortium of academic institutions and we encourage others to consider joining the group of adopting and collaborating sites.

References

[1] Zerhouni EA. A new vision for the national institutes of health. J Biomed Biotechnol 2003;2003(3):159–60.
[2] US Department of Health and Human Services NIoHNCfRR. 2004-2008 Strategic Plan: Challenges and Critical Choices. NIH Publication No.04-5480. 1-1-2004. Ref Type: Report.
[3] Nadkarni PM, Cheung KH. SQLGEN: a framework for rapid client-server database application development. Comput Biomed Res 1995;28(6):479–99.
[4] Nadkarni PM, Brandt CM, Marenco L. WebEAV: automatic metadata-driven generation of web interfaces to entity-attribute-value databases. J Am Med Inform Assoc 2000;7(4):343–56.
[5] Fraternali P, Paolini P. Model-driven development of Web applications: the AutoWeb system. ACM Trans Inf Syst 2000;18(4):323–82.
[6] PHP Hypertext Preprocessor. http://www.php.net/. 2007. Ref Type: Electronic Citation.
[7] MySQL Database Engine. http://www.mysql.com/. 2007. Ref Type: Electronic Citation.
[8] Zubatch M. Value of Hosted Clinical Data Environments. Bio-IT World. 4-14-2006. Ref Type: Magazine Article.
[9] Weber BA, Yarandi H, Rowe MA, Weber JP. A comparison study: paper-based versus web-based data collection and management. Appl Nurs Res 2005;18(3):182–5.