A Statistician's Role in the Utilization of REDCap in an Academic Research Setting

Hannah L. Palac[1,2]
[1]Northwestern University, 680 N. Lake Shore Drive, Suite 1400, Chicago, IL 60611
[2]AbbVie, Inc., 1 N. Waukegan Road, North Chicago, IL 60064

Abstract
High quality study data that is appropriately structured for statistical analyses requires considerable coordination and management from across the entire study team, including data scientists, statisticians, and coordinators. REDCap (Research Electronic Data Capture) is a consortium-based web application used to build research databases quickly and in compliance with institutional standards for security. Some advantages to REDCap include empowering researchers to create case report forms without the need for complex programming, internal user management, audit trails, randomization implementation, and being free to use at many non-profit institutions.

While the REDCap platform requires minimal background knowledge or technical experience, consultation with a statistician is highly recommend to ensure sound data collection and lay the groundwork for smooth data cleaning and statistical analysis post-trial. This paper outlines the role of a statistician in collaborating with researchers to ensure a solid data capture workflow from within the REDCap application.

Key Words: REDCap, data management, data collection

1. Background

High quality data capture and management is crucial to the success of clinical trials. Data that is appropriately structured for statistical analysis often requires considerable coordination involving several members of a research team, however data management platforms are often costly and may require substantial training due to steep learning curves. REDCap (Research Electronic Data Capture) is a user-friendly, web-based program used to build research databases quickly and in compliance with institution security standards.[1] REDCap empowers researchers to create case report forms and surveys without complex programming, facilitates data entry for multicenter studies, can be used in 21 CFR Part 11, FISMA, and HIPAA-compliant environments, can implement randomization models, and is free to use at many non-commercial institutions. Since its development at Vanderbilt in 2004, the number of REDCap users has grown exponentially, reaching over 500,000 users in over 2,500 institutions in 116 countries.

1.1 Support and Infrastructure

Although REDCap is free to use, institutions must comply with infrastructure requirements and execute a valid end-user license agreement with Vanderbilt University. Technical requirements include a PHP web server, MySQL database server, and SMTP e-mail server. A detailed description of requirements and license terms are further outlined

on the REDCap website (https://projectredcap.org/). Once an institution executes a license agreement and joins the REDCap Consortium, the platform is installed on a local web server by institutional IT staff. End-users at the institution then access REDCap over the web in a web browser or through the REDCap Mobile App.

Although there may be some institutional differences in REDCap support models, there are generally two roles that a user can be: an administrator or end-user. REDCap administrators, also known as superusers, are largely information technology (IT) or informatics professionals and provide oversight for the REDCap system as a whole at a given institution. Administrators support REDCap installation, hosting, server maintenance, software upgrades, manage user privileges at an institutional level, and have may have additional capabilities that the typically end-user does not have, such as advanced customizations to the interface and making critical changes after a project is in Production (collecting real study data). Additionally, administrators will typically be the personnel to participate in the REDCap Consortium, which includes a network of institutional partners who utilize REDCap in various ways and provides expansive REDCap support tools. End-users are responsible for building REDCap projects, project testing, data entry, data collection workflow, and setting up project-specific user privileges. An end-user can include data managers, principal investigators (PIs), research coordinators and assistants, statisticians, and other research team members. Both administrators and end-users may take part in REDCap training and troubleshooting.

To facilitate high quality data capture and ensure that data is suitably structured for eventual statistical analyses, it is highly recommended that a statistician be consulted. Further, a statistician can offer valuable skills and knowledge for implementing advanced features.

1.2 Limitations

Like any software, the REDCap software has its limitations such that it may not be suitable for capturing non-relational data, such as multimedia or high-resolution imaging data, does not contain a comprehensive trial management suite with billing capabilities, and may require workarounds to meet study-specific data collection needs. Further, REDCap's flexibility and ease of use may lead to misuse without proper education, communication, and oversight.

1.3 Objective

The objective of this paper is to highlight the crucial role a statistician can have in the proper utilization of the REDCap platform for data capture and management and also in the implementation of advanced REDCap applications. While some specific features are detailed, this paper is not intended to be a tutorial or comprehensive user-guide for the REDCap system. Specifically, this paper will highlight three domains: 1) guidance and consultation; 2) randomization implementation; and 3) use of the REDCap application programming interface (API). Although the experience noted here comes from a statistician's experience in an academic research setting, this guidance is applicable across institutional settings.

2. A Statistician's Role

2.1 Guidance and Consultation

Statisticians and biostatisticians play a powerful role in biomedical research both in academic and non-academic settings. Many statisticians already offer consultation and

collaboration on several aspects of study planning and analysis, including database design and management, and are in a position to build advising on items specific to the REDCap platform into their current work and collaborations. This includes guidance on choosing appropriate field types, variable names, considerations for longitudinal data collection, and consistent numeric coding schemes.

2.2.1 Measurement

Collaborative decisions on *what* to measure to aid in answering a given research question and advising on *how* to collect and record a given measurement play a major role in statistical consultation. Decisions on whether data are best collected as character or numeric variables and advising on formatting standards, such as date formats, parallels with the selection of appropriate field types in REDCap. There are currently (v7.4.6) 13 supported field types in REDCap:

- **Text Box:** Single line text box used to store short text and numbers with options to apply data formatting and validation standards, such as feasible ranges for number entries and date/time formats (among others).
- **Notes Box:** Multiple line text box used to store free-text paragraph entries.
- **Calculated Field:** Text box used to perform real-time calculations automatically (e.g. BMI calculation from height and weight entries).
- **Multiple Choice – Drop-down List:** Single answer dropdown menu displaying multiple choice options. Multiple choice options are assigned a numerical code that will appear as numeric entries when raw data is exported (e.g. 1, Male | 2, Female).
- **Multiple Choice – Radio Buttons:** Single answer radio button displaying multiple choice options. Multiple choice options are assigned a numerical code that will appear as numeric entries when raw data is exported (e.g. 1, Male | 2, Female).
- **Checkbox:** Multiple answer checkbox displaying multiple choice options. Multiple choice options are assigned a numerical code that will appended to variable names when raw data is exported (e.g. race___1, race___2, etc.).
- **Yes – No:** Single answer radio button with pre-coded multiple choice options - 1, Yes | 0, No.
- **True – False:** Single answer radio button with pre-coded multiple choice options - 1, True | 0, False.
- **Signature:** Allows users to draw signature with mouse or figure.
- **File Upload:** Allows users to upload files to a data entry form that can be downloaded to view.
- **Slider / Visual Analog Scale:** Displays slider bar coded as 0-100 with options to customize anchor labels.
- **Descriptive Text:** Allows text, images, videos, or audio files to be viewed in-line or as an attachment. No data is captured in descriptive text fields.
- **Begin New Section/Section Header:** Applies headers and page breaks for aesthetic purposes only. No data is captured in section headers.

Statisticians can ensure that data is being collected using appropriate fields and in the necessary format to perform planned statistical analyses and can also counsel on important considerations that may impact an analysis, such as advising against the use of free text fields, using checkboxes sparingly, and using caution with calculations. In most cases, data entered as free text must be adjudicated and coded after data entry is complete, creating opportunity for error and also taking a substantial amount of time. By

default, checkboxes that are not selected are coded as 0 (unchecked), even if no other data has been collected which may have unintended consequences when making assumptions about missing data and overall data management for a study. It is often advised that raw study data contain only entered data and do not contain calculations; however, there are situations where calculations may be appropriate to be performed during data collection and statisticians are poised to write and review complex equations, as deemed appropriate.

2.2.2 Database design

Many of the perils of data collection using spreadsheets are mitigated with the utilization of REDCap. Some specific areas with guidance for statistician involvement are listed below:

Variable names – The REDCap system does not allow variable names to contain spaces or special characters, except for underscores, converts all variable names to lower case, and gives the user a warning when a variable name is more than 26 characters. That being said, it is important for a statistician to advise on the significance of short, yet meaningful variable names.

Branching/skip logic – Branching logic provides specific conditions to show and/or hide fields based on previous response entries, improving monitoring of missing values and overall data collection workflow. While REDCap does not require a programming background, a solid understanding if applying 'if then' and 'and/or' logic, as often done in statistical programming, is very useful in writing complex branching logic syntax.

Numeric codes and code consistency – REDCap automatically forces the user to apply numeric codes to data labels when using multiple choice field types. Statisticians should ensure that multiple choice options and corresponding numeric codes are consistent throughout a database. For example, it is recommended to use a consistent code when working with "Other" responses or "Unknown" responses for ease of data review when starting an analysis.

Noting missing values – It is highly recommended to use a consistent code for missing values. While this can be accomplished in REDCap by using a consistent numeric code for multiple choice options or entering a specified number in a text box, it may be preferred to use some of the built-in data quality tools, such as the Field Comment Log. The Field Comment Log allows a user to leave a field blank and enter a comment regarding the reason for missing data. The log can then be exported and queried as needed.

2.2 Randomization

The REDCap randomization module allows researchers to implement a pre-specified allocation table and randomly assign participants to treatment groups. To use the randomization module, the first step is to enable the module via the Project Setup Checklist while still in Development mode. The module is enabled by selecting the 'Enable' button next to 'Randomization module' in the 'Enable optional modules and customizations step in the Project Setup Checklist. Once the module is enabled, a separate step will appear in the Project Setup Checklist for the randomization module to be defined.

2.2.1 Randomization setup

The randomization model is defined in three main steps. The first step consists of defining if stratified randomization and/or randomization by group/site will be used and also selecting the field/variable that has been previously defined as the randomization field in the Online Designer or Data Dictionary. When setting up this field, it must be a multiple choice field type. Once this variable has been selected as the randomization field, it will be replaced with a Randomization Button on the data collection form. If stratified randomization is being used, the fields/variables corresponding to the strata must also be reported in this step.

The second step consists of downloading the allocation table template. While allocation tables can be generated in any external program, tables must then be converted to a CSV file prior to upload into REDCap. The template therefore provides a helpful guide for formatting and coding specifications. REDCap will subsequently assign treatment allocations according to this table.

Lastly, the allocation table is uploaded to REDCap. The module is setup such that two separate tables should be uploaded: one while in Development mode for testing and a final table for use in Production status for "real" data collection. The application will check for possible errors, such as using an incorrect variable name for a strata or randomization field or extra or erroneous codes not previously defined in the metadata. Once the project is moved to Production status, tables become locked and unable to be modified. It is highly recommended that tables include more assignments than thought to be needed to account for attrition and enrolling of additional subjects.

2.2.2 Response-adaptive randomization

While the "out of the box" application for randomization within REDCap is setup for fixed allocation schemes, sophisticated use of the REDCap API to integrate complex allocation algorithms such as minimization have been reported.[2] The approach described by Walker and Milne at the 38th Annual Meeting of the Society For Clinical Trials and the 4th International Clinical Trials Methodology Conference utilized the data trigger mechanism and supported the implementation of response-adaptive minimization within REDCap by linking two separate systems while still appearing as a single system to data enterers.

2.2.3 Statistician involvement

It is highly recommended that a statistician performed all setup and management of the REDCap randomization module. Statisticians are poised to understand the manner in which a chosen randomization method is implemented and appreciate the careful attention that this must receive and the importance of maintaining the blind.

2.3 REDCap API

The REDCap API is a powerful application for interfacing with the REDCap platform. It allows external applications, such as statistical software programs or other programming tools, to connect to REDCap and retrieve and or modify data. It is especially useful for performing automated data exports and transferring data between REDCap projects. A statistician may have valuable programming expertise to allow for complex data manipulation, such as aggregating data for reports or monitoring subjects safety, utilizing adaptive randomization via external programs, and integrating REDCap databases with external dashboards, such as R Shiny or Excel. API utilization can dramatically improve

the flexibility of a REDCap database and subsequently improve quality control processes, avoid errors, and significantly decrease time spent on manual tasks.

3. Conclusions

REDCap is a valuable tool for data collection and management for clinical trials run in non-commercial settings. While the REDCap platform requires minimal background knowledge or technical experience, statistical collaboration is critical to optimal utilization of the REDCap platform for data capture to ensure data consistency and lay the groundwork for smooth data cleaning and statistical analysis. Specifically, statisticians can offer helpful expertise and consultation in the use of advanced applications, such as the randomization module and the REDCap API.

Acknowledgements

REDCap is supported at the Feinberg School of Medicine by the Northwestern University Clinical and Translational Science (NUCATS) Institute. Research reported in this publication was supported, in part, by the National Institutes of Health National Center for Advancing Translational Sciences, Grant Number UL1TR001422. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Hannah Palac is currently employed by AbbVie, Inc. The abstract for the 2017 JSM presentation corresponding to this proceedings paper was submitted while she was employed by Northwestern University.

References

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. April 2009;42(2): 377-381.

Walker A, Milne G. Taking the REDCap data management tool to the next step. Reducing complexity in study design by smoother integration of external randomization services thus providing a more streamlined experience for users. Trials. 2017;18 (Suppl 1): O79.